
Algorithmic Methods for Systems Biology of Herpes-viral microRNAs

Florian Erhard



München 2013

Algorithmic Methods for Systems Biology of Herpes-viral microRNAs

Florian Erhard

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Florian Erhard
geboren in Schongau

München, den 18.10.2013

Erstgutachter: Prof. Dr. Ralf Zimmer

Zweitgutachter: Prof. Dr. Rolf Backofen

Tag der mündlichen Prüfung: 27.03.2014

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Erhard, Florian

Name, Vorname

München, 18.10.2013

Ort, Datum

Unterschrift Doktorand/in

Contents

Summary	xv
Zusammenfassung	xvii
1 Introduction	1
1.1 Systems biology	1
1.1.1 A typical systems biology workflow	2
1.1.2 Bioinformatics in systems biology	2
1.2 MicroRNAs	7
1.2.1 Discovery of microRNAs in <i>C. elegans</i>	7
1.2.2 MicroRNA biogenesis	8
1.2.3 Regulatory mechanisms of microRNAs	8
1.2.4 Biological function of microRNAs	10
1.2.5 Bioinformatics for microRNAs	11
1.3 Herpes viruses	14
1.3.1 Phylogeny	14
1.3.2 Life cycle, symptoms and prevalence	15
1.3.3 Viral microRNAs	16
2 Datasets	19
2.1 Primer on experimental techniques	19
2.1.1 Microarrays, metabolic labeling and RIP-Chip	19
2.1.2 Sequencing, sRNA-seq and PAR-CLIP	21
2.1.3 LC-MS/MS and SILAC	22
2.2 Cell lines and available datasets	23
2.2.1 KSHV related cell lines	24
2.2.2 EBV related cell lines	26
2.2.3 VZV related cell lines	26
3 Classification of ncRNAs using position and size information in deep sequencing data	29
3.1 Abstract	30
3.1.1 Motivation	30

3.1.2	Results	30
3.1.3	Availability	30
3.2	Introduction	30
3.3	Approach	32
3.4	Methods	34
3.5	Results	37
3.6	Discussion	40
3.7	Conclusion and Outlook	42
4	PARma: identification of microRNA target sites in AGO-PAR-CLIP data	43
4.1	Abstract	44
4.2	Introduction	44
4.3	Results	46
4.3.1	PARma overview	46
4.3.2	Cluster detection	48
4.3.3	Generative model	52
4.3.4	KmerExplain	53
4.3.5	Seed activities	53
4.3.6	Inferred models	55
4.3.7	Evaluation using differential PAR-CLIP	56
4.3.8	Validation against RIP-Chip data	59
4.4	Discussion	60
4.4.1	PAR-CLIP clusters	60
4.4.2	PARma	61
4.4.3	Comparison to PARalyzer	62
4.4.4	Differential PAR-CLIP	63
4.5	Methods	64
4.5.1	Data	64
4.5.2	Raw data processing and cluster definition	64
4.5.3	PARma	66
4.6	Software availability	69
5	RIP-chip enrichment analysis	71
5.1	Abstract	72
5.1.1	Motivation	72
5.1.2	Results	72
5.1.3	Availability	72
5.2	Introduction	72
5.3	Methods	75
5.3.1	Data processing	75
5.3.2	Mixture model fitting	75
5.3.3	PCA	76
5.4	Results	77

5.4.1	Select relevant genes	77
5.4.2	Determining microRNA targets	79
5.4.3	Taking replicates into account	80
5.4.4	Determining differential microRNA targets	82
5.5	Discussion	83
5.6	Conclusion	85
6	Widespread context-dependency of microRNA-mediated regulation	87
6.1	Abstract	89
6.2	Introduction	89
6.3	Results	91
6.3.1	Differential analysis of PAR-CLIP data	91
6.3.2	Technical bias	94
6.3.3	Context-dependent target sites of KSHV microRNAs	96
6.3.4	Context-dependent target sites of cellular microRNAs	103
6.3.5	mRNA levels and flanking sequence motifs explain context-dependent microRNA/target interactions	106
6.3.6	Context-dependent target sites are less conserved than constitutive sites	110
6.4	Discussion	111
6.4.1	Contributors to the cellular context	111
6.4.2	Other contributors	112
6.4.3	Functional considerations of context-dependent regulation	113
6.4.4	Consequences of context-dependency	114
6.5	Methods	115
6.5.1	Cell lines	115
6.5.2	PAR-CLIP and sequencing	115
6.5.3	SILAC-based proteomics	116
6.5.4	RIP-Chip analysis	116
6.5.5	RNA half-life measurements by 4sU-tagging	116
6.5.6	PARma	116
6.5.7	Correcting for sampling noise	117
7	Detection of outlier peptides	121
7.1	Abstract	122
7.2	Introduction	122
7.3	Materials and methods	125
7.3.1	Data processing	125
7.3.2	Detecting outlier peptides	125
7.3.3	In-silico data generation	126
7.4	Results and Discussion	127
7.4.1	Test on in-silico generated data	128
7.4.2	Outlier peptides in real data	129
7.4.3	Discussion	135

7.5	Conclusion	136
8	FERN - Stochastic Simulation and Evaluation of Reaction Networks	137
8.1	Abstract	138
8.2	Background	138
8.2.1	Petri nets	139
8.2.2	Stochastic chemical kinetics	140
8.2.3	Stochastic simulation methods	142
8.3	Implementation	145
8.3.1	Other implementations	146
8.3.2	FERN	146
8.3.3	Implementation details	148
8.3.4	Accuracy and runtime performance of FERN	151
8.4	Using FERN	152
8.4.1	Command line tool	152
8.4.2	Basic usage of FERN	153
8.4.3	Cytoscape plugin for stochastic simulation	154
8.4.4	Simulation of cell growth and division using observers	156
8.5	Discussion	157
9	Conclusion and outlook	159
	Bibliography	163
	List of abbreviations	189
	Acknowledgements	191

List of Figures

1.1	Systems biology workflow	3
1.2	Phylogeny of herpes viruses	15
2.1	Proteomics data for DG75-eGFP,DG75-10/12 and BCBL1	25
2.2	Read length distribution of the VZV sRNA-seq experiments	27
3.1	Typical position and length dependent pattern matrices	33
3.2	Freeshift alignment of hsa-mir-99b and hsa-mir-185	37
3.3	Inner and outer score distributions for microRNAs	38
3.4	Evaluation for the scoring system ($H^{2,1}$, sum-min, -0.01 , -0.005 , freeshift) . . .	40
4.1	PARma overview	47
4.2	PAR-CLIP data viewer	49
4.3	Illustration of the PARma procedure	50
4.4	Overlapping PAR-CLIP clusters	51
4.5	Correlation of microRNA expression to the number of assigned clusters	54
4.6	PARma model for replicate A of the DG75 experiment	55
4.7	Model scores for the cluster in Figure 4.2a	57
4.8	Evaluation using differential PAR-CLIP	58
4.9	Validation against RIP-Chip	60
5.1	Measurement distributions for our Ago2 RIP-chip experiment	73
5.2	Selecting expressed genes	78
5.3	Background subtraction is necessary	79
5.4	Computed FDRs are valid	81
5.5	Average number of HITS-Clip target sites per target/non-target gene	82
5.6	Differing IP efficiencies require normalization before computing summary values	83
5.7	Differing IP efficiencies require normalization before computing differential targets	84
6.1	Validation of PAR-CLIP experiments	92
6.2	Comparison of PAR-CLIP experiments with available datasets	93
6.3	Comparison of PAR-CLIP datasets	94
6.4	Correlations of PAR-CLIP cluster quantifications	95
6.5	miR-K12-4-3p heatmap	97

6.6	Context dependent target sites of KSHV microRNAs	98
6.7	KSHV PAR-CLIP targets in RIP-Chip data	99
6.8	KSHV PAR-CLIP targets in mRNA half-life data	101
6.9	KSHV PAR-CLIP targets in expression data	102
6.10	Context-dependent target interactions of human microRNAs	103
6.11	Cellular PAR-CLIP targets in RIP-Chip and mRNA half-life data	104
6.12	Cellular PAR-CLIP targets in expression data	105
6.13	Comparison of mRNA fold changes to PAR-CLIP read count fold changes	107
6.14	Role of sequence motifs for context-dependent target sites	108
6.15	Motif randomization results	109
6.16	Conservation of target sites	110
6.17	PAR-CLIP read count correlation with expression	117
6.18	Conditional gamma distribution fit	119
7.1	Example MS peptide quantifications for a gene with several isoforms	124
7.2	Number of measurements per peptide	128
7.3	Evaluation on in-silico generated data	130
7.4	Heteroscedastic ANOVA applied to the experimental data	131
7.5	Evidence for misidentifications in outlier peptides	132
7.6	Evidence for misquantifications in outlier peptides	133
7.7	Evidence for saturation in our dataset	134
8.1	A Petri net and the firing of a transition.	142
8.2	Flow of one simulation step	143
8.3	Software design	147
8.4	Trajectories of the EGF signalling pathway	153
8.5	EGF signalling pathway loaded into Cytoscape	155
8.6	Average results of 1,000 simulations for LacZ	157

List of Tables

1.1	Experimentally discovered microRNAs of human herpes viruses	17
2.1	Datasets for the cell lines DG75-eGFP, DG75-10/12 and BCBL1	24
3.1	Annotations from mirBase, gtRNAdb, Ensembl and RefSeq	36
6.1	Identified motifs by MERCI	106
8.1	Mass action propensity functions for basic reactions	141

Summary

Recent technological advances have made it possible to measure various parameters of biological processes in a genome-wide manner. While traditional molecular biology focusses on individual processes using targeted experiments (reductionistic approach), the field of systems biology utilizes high-throughput experiments to determine the state of a complete system such as a cell at once (holistic approach). Systems biology is not only carried out in wet-lab, but for the most part also requires tailored computational methods. High-throughput experiments are able to produce massive amounts of data, that are often too complex for a human to comprehend directly, that are affected by substantial noise, i.e. random measurement variation, and that are often subject to considerable bias, i.e. systematic deviations of the measurement from the truth. Thus, computer science and statistical methods are necessary for a proper analysis of raw data from such large-scale experiments.

The goal of systems biology is to understand a whole system such as a cell in a quantitative manner. Thus, the computational part does not end with analyzing raw data but also involves visualization, statistical analyses, integration and interpretation. One example for these four computational tasks is as follows: Processes in biological systems are often modeled as networks, for instance, gene regulatory networks (GRNs) that represent the interactions of transcription factors (TFs) and their target genes. Experiments can provide both, the identity and wiring of all constituent parts of the network as well as parameters that allow to describe the processes in the system in a quantitative manner. A network provides a straight-forward way to visualize the state and processes of a whole system, its statistical analysis can reveal interesting properties of biological systems, it is able to integrate several datasets from various experiments and simulations of the network can aid to interpret the data.

In recent years, microRNAs emerged as important contributors to gene regulation in eukaryotes, breaking the traditional dogma of molecular biology, where DNA is transcribed to RNA which is subsequently translated into proteins. MicroRNAs are small RNAs that are not translated but functional as RNAs: They are able to target specific messenger RNAs (mRNA) and typically lead to their downregulation. Thus, in addition to TFs, microRNAs also play important roles in GRNs. Interestingly, not only animal genomes including the human genome encode microRNAs, but microRNAs are also encoded by several pathogens such as viruses.

In this work I developed several computational systems biology methods and applied them to high-throughput experimental data in the context of a project about herpes viral microRNAs. Three methods, ALPS, PARma and REA, are designed for the analysis of certain types of raw data, namely short RNA-seq, PAR-CLIP and RIP-Chip data, respectively. All of these

experiments are widely used and my methods are publicly available on the internet and can be utilized by the research community to analyze new datasets. For these methods I developed non-trivial statistical methods (e.g. the EM algorithm `kmerExplain` in `PARma`) and implemented and adapted algorithms from traditional computer science and bioinformatics (e.g. alignment of pattern matrices in `ALPS`).

I applied these novel methods to data measured by our cooperation partners in the herpes virus project. I.a., I discovered and investigated an important aspect of microRNA-mediated regulation: MicroRNAs recognize their targets in a context-dependent manner. The widespread impact of context on regulation is widely accepted for transcriptional regulation, and only few examples are known for microRNA-mediated regulation. By integrating various herpes-related datasets, I could show that context-dependency is not restricted to few examples but is a widespread feature in post-transcriptional regulation mediated by microRNAs. Importantly, this is true for both, for human host microRNAs as well as for viral microRNAs.

Furthermore, I considered additional aspects in the data measured in the context of the herpes virus project: Alternative splicing has been shown to be a major contributor to protein diversity. Splicing is tightly regulated and possibly important in virus infection. Mass spectrometry is able to measure peptides quantitatively genome-wide in high-throughput. However, no method was available to detect splicing patterns in mass spectrometry data, which was one of the datasets that has been measured in the project. Thus, I investigated whether mass spectrometry offers the opportunity to identify cases of differential splicing in large-scale.

Finally, I also focussed on networks in systems biology, especially on their simulation. To be able to simulate networks for the prediction of the behavior of systems is one of the central goals in computational systems biology. In my diploma thesis, I developed a comprehensive modeling platform (`PNMA`, the Petri net modeling application), that is able to simulate biological systems in various ways. For highly detailed simulations, I further developed `FERN`, a framework for stochastic simulation that is not only integrated in `PNMA`, but also available stand-alone or as plugins for the widely used software tools `Cytoscape` or `CellDesigner`.

In systems biology, the major bottleneck is computational analysis, not the generation of data. Experiments become cheaper every year and the throughput and diversity of data increases accordingly. Thus, developing new methods and usable software tools is essential for further progress. The methods I have developed in this work are a step into this direction but it is apparent, that more effort must be devoted to keep up with the massive amounts of data that is being produced and will be produced in the future.

Zusammenfassung

Der technische Fortschritt in den letzten Jahren hat ermöglicht, dass vielerlei Parameter von biologischen Prozessen genomweit gemessen werden können. Während die traditionelle Molekularbiologie sich mit Hilfe gezielter Experimente auf individuelle Prozesse konzentriert (reduktionistischer Ansatz), verwendet das Feld der Systembiologie Hochdurchsatz-Experimente um den Zustand eines vollständigen Systems wie einer Zelle auf einmal zu bestimmen (holistischer Ansatz). Dabei besteht Systembiologie nicht nur aus Laborarbeit, sondern benötigt zu einem großen Teil auch speziell zurechtgeschnittene computergestützte Methoden. Hochdurchsatz-Experimente können riesige Mengen an Daten produzieren, welche oft zu komplex sind um von einem Menschen direkt verstanden zu werden, welche beeinträchtigt sind von substantiellem Rauschen, das heißt zufälliger Messvariation, und welche oft beträchtlichem Bias unterliegen, also systematischen Abweichungen der Messungen von der tatsächlichen Größe. Daher sind informatische und statistische Methoden notwendig für eine geeignete Analyse der Rohdaten eines groß angelegten systembiologischen Experiments.

Das Ziel der Systembiologie ist ein ganzes System wie eine Zelle in quantitativer Weise zu verstehen. Daher endet der computergestützte Teil nicht mit der Analyse der Rohdaten, sondern beinhaltet ebenfalls Visualisierung, statistische Analyse, Integration und Interpretation. Ein Beispiel dieser vier rechnergestützten Aufgaben ist wie folgt: Prozesse in biologischen Systemen werden oft in Netzwerken modelliert. Zum Beispiel werden in genregulatorischen Netzwerken (GRNs) die Interaktionen zwischen Transkriptionsfaktoren (TFs) und deren Zielgenen repräsentiert. Mit Experimenten kann man sowohl die Identität und die Vernetzung aller Bestandteile des Netzwerkes messen, wie auch die Parameter, mit denen man die Prozesse des Systems in quantitativer Weise beschreiben kann. Mit Hilfe eines Netzwerkes kann man auf einfache und direkte Weise den Zustand und die Prozesse eines ganzen Systems visualisieren, die statistische Analyse des Netzwerkes kann interessante Eigenschaften eines biologischen Systems aufdecken, es bietet die Möglichkeit, verschiedene experimentelle Daten zu integrieren und seine Simulation kann bei der Interpretation der Daten helfen.

Erst vor wenigen Jahren stellte sich heraus, dass sogenannte microRNAs die Genregulation in Eukaryonten maßgeblich beeinflussen. Das steht im Widerspruch zum traditionellen Dogma der Molekularbiologie, bei dem die genetische Information aus der DNA in RNA transkribiert wird, welche anschließend in Proteine translatiert wird. MicroRNAs hingegen sind kurze RNAs, welche nicht translatiert werden, sondern als RNAs funktional sind. Sie können spezifische messenger RNAs (mRNAs) binden und führen dann typischerweise zu deren Inhibition. Zusätzlich zu Transkriptionsfaktoren spielen also microRNAs eine wichtige Rolle

in GRNs. Interessanterweise enkodieren nicht nur tierische Genome, das menschliche Genom eingeschlossen, microRNAs, sondern viele Pathogene wie Viren exprimieren ihre eigenen microRNAs in infizierten Wirtszellen.

In dieser Arbeit habe ich mehrere computergestützte Methoden für die Anwendung in der Systembiologie entwickelt und auf Hochdurchsatz-Daten angewendet, die im Kontext eines Projektes über herpesvirale microRNAs vermessen wurden. Drei Methoden, ALPS, PARma und REA, habe ich für die Analyse von bestimmten Typen von Rohdaten entworfen, nämlich jeweils short RNA-seq, PAR-CLIP und RIP-Chip. All diese Experimente sind weit verbreitet im Einsatz und meine Methoden sind im Internet öffentlich verfügbar und können von der Forschungsgemeinschaft zur Analyse der Rohdaten der jeweiligen Experimente verwendet werden. Für diese Methoden entwickelte ich nicht-triviale statistische Methoden (z.B. den EM Algorithmus kmerExplain in PARma) und implementierte und adaptierte Algorithmen aus der traditionellen Informatik wie auch aus der Bioinformatik (z.B. Sequenzalignment der Mustermatrizen in ALPS).

Ich wendete diese neuen Methoden auf Daten an, die von unseren Kooperationspartner im Herpesviren Projekt gemessen wurden. Dabei entdeckte und erforschte ich unter anderem einen wichtigen Aspekt der Regulation durch microRNAs: MicroRNAs erkennen ihre Targets in kontext-abhängiger Weise. Die weitverbreiteten Auswirkungen von Kontext ist weithin akzeptiert für transkriptionelle Regulation und es sind nur wenige Beispiele von kontext-spezifischer microRNA gesteuerte Regulation bekannt. Indem ich mehrere Herpes-relevante Datensätze integriert analysiert habe, konnte ich zeigen, dass Kontext-Abhängigkeit nicht nur auf ein paar Beispiele beschränkt ist, sondern dass es ebenfalls ein weitverbreitetes Merkmal der post-transkriptionellen Regulation gesteuert durch microRNAs ist, dass Zielgene kontext-abhängig erkannt werden. Das gilt sowohl für die menschlichen microRNAs der Wirtszelle wie auch für die exogenen viralen microRNAs.

Desweiteren habe ich zusätzliche Aspekte der Daten des Herpesviren-Projektes betrachtet: Es wurde gezeigt, dass alternatives Spleißen maßgeblich zur Diversität von Proteinen beiträgt. Spleißen ist streng reguliert und möglicherweise wichtig bei der Virusinfektion. Massenspektrometrie kann Peptide genomweit in quantitativer Weise messen. Allerdings stand keine Methode zur Verfügung, um Spleiß-Muster in Massenspektrometrie-Daten, wie sie im Projekt gemessen wurden, zu detektieren. Aus diesem Grund habe ich untersucht, ob es mit Massenspektrometrie-Daten möglich ist, Fälle von alternativen Spleißen im großen Umfang zu identifizieren.

Letztendlich habe ich mich auch auf systembiologische Netzwerke und im Speziellen auf deren Simulation konzentriert. Netzwerke simulieren zu können um das Verhalten von Systemen vorherzusagen ist eines der zentralen Ziele der rechnergestützten Systembiologie. Bereits in meiner Diplomarbeit habe dafür ich eine umfassende Modellierplattform (PNMA, the Petri net modelling application) entwickelt. Damit ist es möglich, biologische Systeme auf vielerlei Arten zu simulieren. Für sehr detaillierte Simulationen habe ich dann FERN entwickelt, ein Framework zur stochastischen Simulation, welches nicht nur in PNMA integriert ist, sondern auch als eigenständige Software wie auch also Plugin für die weitverbreiteten Programme Cytoscape und CellDesigner verfügbar ist.

Der Engpass in der Systembiologie ist mehr und mehr die rechnergestützte Analyse der Daten und nicht deren Generierung. Experimente werden jedes Jahr günstiger und der Durchsatz und die Diversität der Daten wächst dementsprechend. Daher ist es für den weiteren wissenschaftlichen Fortschritt essentiell, neue Methoden und benutzbare Softwarepakete zu entwickeln. Die Methoden, die ich in dieser Arbeit entwickelt habe, stellen einen Schritt in diese Richtung dar, aber es ist offensichtlich, dass mehr Anstrengungen aufgewendet werden müssen, um Schritt halten zu können mit den riesigen Mengen an Daten die produziert werden und in der Zukunft noch produziert werden.

Chapter 1

Introduction

1.1 Systems biology

The advent of high-throughput technologies has revolutionized biological research and has heavily contributed to the increasing importance of the field of systems biology [Stelling, 2004; Westerhoff and Palsson, 2004; Kitano, 2002; Ideker et al., 2001; Ideker and Lauffenburger, 2003]. Instead of focussing on single or few biological entities such as genes or proteins, high-throughput methods allow to investigate a complex biological system such as a cell or cell culture in its entirety.

In the top-down approach of systems biology [Ideker and Lauffenburger, 2003], so-called omics experiments are applied, which are based on these high-throughput technologies and measure a certain type of data in a genome-wide manner: These omics experiments include, but are not limited to, genomics measuring the DNA sequence of complete genomes including intra- or inter-species variation [Lander et al., 2001; Venter et al., 2001; Lindblad-Toh et al., 2011; Consortium, 2012b], transcriptomics measuring the identity and quantity of expressed messenger RNAs (mRNAs) and non-coding RNAs (ncRNAs) [Schena et al., 1995; Mortazavi et al., 2008], proteomics measuring the expression and modifications of proteins [Ong and Mann, 2005; Cox and Mann, 2007], metabolomics measuring concentrations of metabolites [Nicholson and Wilson, 2003; German et al., 2005], interactomics measuring all interactions between molecules [Schwikowski et al., 2000; Rual et al., 2005] and many more.

The goal of systems biology is to integrate all these omics measurements in order to understand and characterize processes that are important in the system under consideration in a holistic manner [Sauer et al., 2007; Noble, 2008]. Naturally, due to the massive amount of data from such experiments, bioinformatics plays an irreplaceable role in systems biology [Scholz et al., 2012; Likić et al., 2010].

In this work, I developed computational methods for the analysis of high-throughput experiments and applied them to data measured in the context of a project about herpes viruses with a focus on post-transcriptional regulation by microRNAs. In this first chapter, I give a pragmatic overview of the methods by classifying them into a typical workflow in systems biology and, after an introduction of microRNAs, I describe the impact of the proposed approaches on analyzing

microRNA related experiments in general. A brief introduction of herpes virus biology concludes this chapter. In the second chapter, I briefly describe the experimental techniques and datasets used in this work, followed by an overview about how all developed methods have been applied to the data. Then, I describe all methods and results of the analyses in detail in the subsequent chapters. I conclude this work with an outlook and future developments.

1.1.1 A typical systems biology workflow

The output of a high-throughput experiments is a certain type of data, which by itself is not to be mistaken for information about or understanding of a system. Bioinformatics is needed to extract information and, further on, to interpret the data in order to understand the system. Making the distinction between information and understanding, the role of bioinformatics is twofold: First, the raw data coming from a high-throughput experiment must be analyzed. For instance, data may consist of short sequencing reads or fluorescence intensities from a microarray. Computational methods must be applied to extract useful information, e.g. an expression value per gene in a transcriptomics study, and often these values can be presented in a tabular format. Such tables often contain thousands of rows that represent entities such as genes and dozens of columns representing different pieces of information or different statistics about the data.

Due to the complexity of this massive amount of information, it is often impossible to quickly come to an understanding of the system. This is the second task for bioinformatics: Provide tools and methods to interpret such per-entity values in the context of the system. Such tools and methods may be as simple as overrepresentation analysis [Breitling et al., 2004] or gene set enrichment [Subramanian et al., 2005], or they may make use of more sophisticated methods integrating multiple datasets or existing knowledge [Consortium, 2012b; Gerstein et al., 2012]. In systems biology, such advanced methods are of uttermost importance since the goal is to understand a complex biological system as a whole with all its interdependencies and not only to investigate individual components.

Thus, a typical workflow in top-down systems biology can be subdivided into four steps (see also Figure 1.1): First, a biological system, for instance a cell culture, is subjected to several sample preparation steps, e.g. to isolate mRNA in form of cDNA. Then, the actual high-throughput measurement takes place, for instance RNA-seq, where the sequence of millions of isolated mRNA fragments is determined. After this second step, further work takes place in front of a computer instead of in the laboratory: As indicated above, the raw data must be analyzed, for instance by mapping the RNA-seq reads to known genes and computing per-gene expression values by counting sequenced reads. Finally, these per-gene values can be further analyzed and interpreted depending on the intention of the study. Of course, this workflow makes no claim to be complete and depending on the nature and goal of a study, further wet-lab validations or high-throughput measurements may be necessary [Mortazavi et al., 2008].

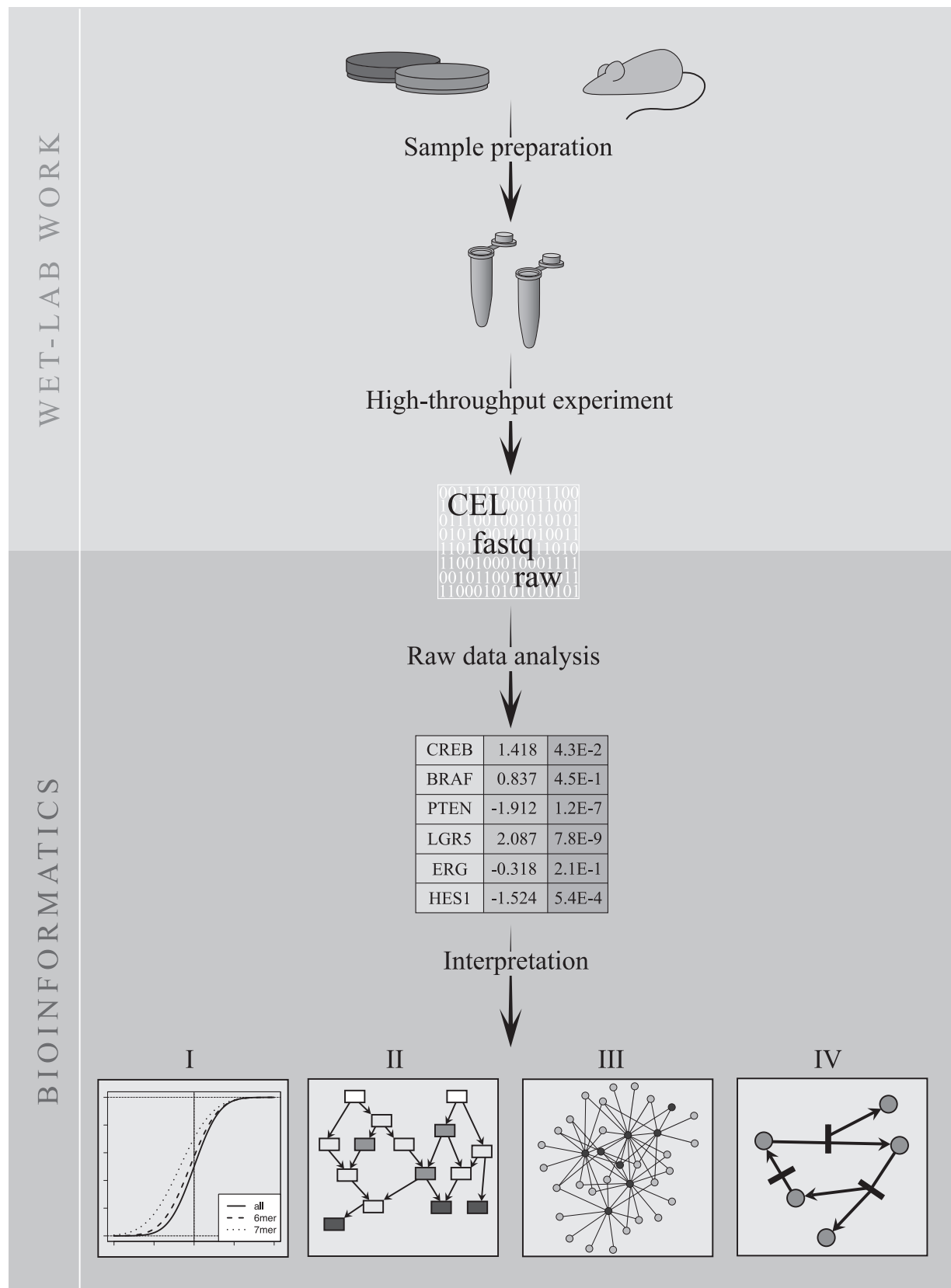
1.1.2 Bioinformatics in systems biology

There are several issues associated with both major computational steps, raw data analysis and downstream interpretation, which I have approached in this work.

First, while most omics experiments are based on common high-throughput technologies, there is a huge variety of differing sample preparation steps. Thus, even if the raw data may be of the same kind for various experiments, the demands on the analysis methods may be highly diverse. For instance, currently a multitude of omics experiments are based on next generation sequencing (NGS), e.g. DNA resequencing in genomics [Consortium, 2012a], RNA-seq in transcriptomics [Mortazavi et al., 2008] or CLIP-seq in interactomics [Chi et al., 2009; König et al., 2010; Hafner et al., 2010]. For all these experiments, the raw data consists of millions of short sequencing reads, ranging from 35 nucleotides (nt) to more than 100 nt. Thus, the first step of data analysis is similar for most experiments, i.e. to align these reads to a reference sequence such as the genome or transcriptome. However, further data analysis is often highly diverse: DNA resequencing usually has the goal to detect single nucleotide polymorphisms (SNPs) or copy number variations (CNVs) and further analysis therefore includes the detection of significant mismatches or coverage differences with respect to the genome [Consortium, 2012a]. In contrast, RNA-seq is designed to quantify all mRNAs in a sample and, thus, methods must be applied to properly estimate mRNA abundance or identify significantly differentially expressed genes [Mortazavi et al., 2008; Anders and Huber, 2010]. Moreover, for CLIP-seq experiments, RNA is crosslinked to proteins, the protein of interest is isolated using immunoprecipitation (IP) and digested fragments crosslinked to this proteins are sequenced [Chi et al., 2009]. Thus, in order to identify the mRNA binding sites of the protein, locations must be found where many sequencing reads have been aligned to. Furthermore, several modifications have been developed for the CLIP protocol [König et al., 2010; Hafner et al., 2010], which lead to additional characteristics in the data for the identification of bona-fide binding sites. For instance, in PAR-CLIP [Hafner et al., 2010], the usage of the uridine analogue 4-thio-uridine (4sU) and a certain wavelength of the laser used for crosslinking leads to so-called T to C conversions, since crosslinked 4sU is read as a C and not as a T during cDNA synthesis. Thus, even if raw data may be similar across a multitude of experiments such as sequencing experiments, specialized experimental assays such as CLIP-seq require specialized analysis methods and considering special features of variants of such assays, e.g. PAR-CLIP, may yield more or more reliable information. In chapter 4, I describe an analysis method I developed that addresses this important issue. PARma utilizes characteristic features of PAR-CLIP data and thereby outperforms existing approaches in identifying reliable microRNA/target interactions Erhard et al. [2013a].

Second, in some cases, existing methods for one kind of experiments may be directly applicable to another kind of experiment to a certain extent, but they may neglect bias that does not affect

Figure 1.1 (*following page*): Systems biology workflow. Starting from a biological system, either in-vitro or in-vivo, samples are prepared and subjected to a high-throughout platform (Wet-lab work). After that, raw data from the experiment must be analyzed, which often results in a table with biological entities in rows (e.g. genes) and various kinds of data in columns. The final step is to interpret these tables, for instance by statistical analysis (I), functional analysis (II), network analysis (III) or simulation of a network model (IV). Often, interpreting these tables leads to new hypothesis for new experiments, i.e. further wet-lab work.



the type of experiment which the methods have been originally designed for. For instance, RIP-Chip, another interactomics experiment type, utilizes microarrays, which are widely applied to mRNA quantification in transcriptomics. In RIP-Chip, RNA binding proteins (RBPs) are co-immunoprecipitated with their associated mRNAs and the abundance of isolated mRNA is then compared to the total abundance of the same mRNA in full cell lysate or the isolate from a control IP using an unspecific antibody [Tenenbaum et al., 2000]. This is very similar to standard differential expression (DE) analysis using microarrays, where total RNA in one condition is compared to total RNA in another condition Miller and Tang [2009]. While in DE, the ratio between conditions corresponds to the expression fold change, for a RIP-Chip experiment, the ratio expresses the enrichment of an mRNA in the IP fraction. Consequently, in previous studies, transcriptomics data analysis methods have directly been applied to RIP-Chip data [Mukherjee et al., 2009; Hendrickson et al., 2008; Karginov et al., 2007; Stoecklin et al., 2008; Landthaler et al., 2008; Dölken et al., 2010]. However, differing efficiencies of the IP that are generally observed in RIP-Chip experiments introduces severe bias that must be accounted for before further analyzing RIP-Chip data. This is topic of chapter 5 and has been published in Erhard et al. [2013b].

Third, there may be information present in high-throughput data that is not considered by existing analysis tools and awaits uncovering by specialized methods. This information can be subcategorized into (a) additional characteristics of the entities that were originally intended to be measured and (b) features of additional entities that were not in the focus of the original experimental design. An example of additional characteristics is described in chapter 5 and published in Erhard et al. [2013b]: By inspecting the distribution of IP enrichments of all mRNAs, it is possible to test whether an mRNA is significantly targeted by an RBP by computing false discovery rates (FDRs). Importantly, this FDR does not correspond to the reproducibility of the enrichment but assesses whether the magnitude of the enrichment is high enough to speak of a functionally relevant enrichment. Therefore, this is complementary to other methods to compute FDRs that are computed by considering replicate measurements, e.g. by using *t* statistics [Mukherjee et al., 2009] or moderated *t* statistics [Hendrickson et al., 2008], and assess the reproducibility of the enrichment. Thus, the FDR as computed by our method represents an additional property of a set of genes that is hidden in the data and can be uncovered by specialized analysis methods. Furthermore, as indicated above, an experiment may also yield additional data about entities that were not originally intended to be measured by the assay's design. For instance, the goal of small RNA-seq experiments is to profile the expression of a certain kind of ncRNAs, namely microRNAs (see below). However, in such an experiment, not only microRNAs are sequenced, but a variety of other known and unknown ncRNAs or ncRNA fragments. Usually, these are discarded and excluded from further analysis. However, we have shown that relative positions and lengths of these fragments provide information for the important task of classification of ncRNAs [Erhard and Zimmer, 2010], which is described in chapter 3.

Fourth, even if information has been extracted from the experimental data properly, i.e. if proper per-entity values are available, specialized downstream analysis methods may be necessary to answer specialized biological questions. For instance, alternative splicing is one of the key contributors to the diversity of gene products observed for most multi-cellular organisms [Wang and Burge, 2008]. Importantly, the alternative splicing pattern may be highly diverse across

different conditions such as cell types and finding differentially spliced genes is therefore a hot topic in current RNA-seq research [Richard et al., 2010; Trapnell et al., 2013]. However, differential splicing of mRNAs only has functional impact if the corresponding transcripts are translated into proteins. Therefore, we sought to identify differential splicing on protein level using stable isotope labelling of amino acids in cell culture (SILAC) based high-throughput liquid chromatography tandem mass spectrometry (LC-MS/MS) data. In comparison to RNA-seq data, LC-MS/MS data is more sparse, i.e. there are usually only few peptides measured per transcript in comparison to transcripts that are often fully covered by sequencing reads. Thus, we investigated, how and to what extent LC-MS/MS data can nevertheless be used to reliably find cases of differentially spliced genes [Erhard and Zimmer, 2012], which is described in chapter 7. Fifth, the variety of available omics experiments leads to a growing demand of methods to integrate different datasets, which is not a trivial task. For instance, in such experiments, distinct entities may be primarily measured: While most microarrays measure expression of genes by using probesets directed against 3'-untranslated regions (UTRs) (that are often common to all transcript isoforms), RNA-seq can be used to directly measure transcript abundances, CLIP-seq identifies binding sites and SILAC based LC-MS/MS measures fold changes of peptides. It is of great importance to integrate all these measurements properly for many reasons. For instance, since high-throughput experiments are affected by noise, i.e. random measurement errors, and bias, i.e. a systematic deviation from the truth, interpretations based on such data must be validated. This is often done by targeted experiments that are more reliable but not applicable in large-scale, or, alternatively, by additional independent high-throughput experiments. In some cases, considering additional high-throughput experiments may be the better choice to confirm interpretations: In chapter 6 I present our analysis of the widespread context-dependence of microRNA-mediated regulation. This context-dependence has already been shown for few examples by using targeted experiments. However, without large-scale experiments, it is impossible to judge whether these are rare exceptions or if context-dependence is a general feature of microRNA-mediated regulation. Thus, we integrated multiple datasets to confirm this hypothesis of widespread context-dependence of microRNA-mediated regulation [Erhard et al., 2013c], namely RIP-Chip experiments of the microRNA containing RNA induced silencing complex (RISC), AGO-PAR-CLIP, microarray experiments for mRNA steady-state levels and half-lives based on metabolic labeling and SILAC based LC-MS/MS experiments for protein expression levels.

Finally, as indicated above, simple overrepresentation analysis of the set of differentially expressed genes among predefined gene sets [Breitling et al., 2004] or gene set enrichment analysis on continuous data such as all fold changes of mRNAs [Subramanian et al., 2005] may provide a first handle to interpret the information from a high-throughput experiment. However, they fall far too short for a mechanistic understanding of processes or a whole system. Therefore, a widely used approach in computational systems biology is to model a biological system as a network or graph: For instance, the interplay between all contributors of gene regulation is commonly called gene regulatory network (GRN). Here, each node of the network represents a certain gene and a node A is connected to a node B, if the gene product of A regulates the expression of B. If detailed information, e.g. coming from multiple high-throughput datasets, is available about a system, such a network-based model can be simulated [Jaeger et al., 2004;

Segal et al., 2008; Erhard, 2008]. Hence, a valid model should be able to explain or reproduce the data of the system. In my diploma thesis [Erhard, 2008], I started the development of PNMA, the Petri net modeling application, which is a comprehensive modeling tool for systems biology. In chapter 8, I describe a further development of PNMA, the integration of methods for stochastic simulation (FERN, Erhard et al. [2008]).

1.2 MicroRNAs

Only in recent years, it has become apparent that gene regulation is not only carried out by transcription factors (TFs), i.e. during transcription, but that there is also another regulatory layer that controls expression levels of genes post-transcriptionally. MicroRNAs (often also referred to as miRNAs or miRs) are small, 20-24 nt long RNA molecules that have emerged as important mediators of post-transcriptional gene regulation [Bartel, 2004; He and Hannon, 2004; Bartel, 2009; Ghildiyal and Zamore, 2009; Pasquinelli, 2012]. They can be found in all kingdoms of life, most prominently in metazoans [Pasquinelli et al., 2000; Wheeler et al., 2009] and in plants [Jones-Rhoades et al., 2006], but also in viruses [Kincaid and Sullivan, 2012] and bacteria [Zhao et al., 2007]. Additionally, microRNA like molecules have also been identified in fungi [Lee et al., 2010].

1.2.1 Discovery of microRNAs in *C. elegans*

More than 20 years ago, a genetic screen identified a genetic locus called *lin-4* that takes part in the control of larval development in the nematode *Caenorhabditis elegans*. Loss-of-function of *lin-4* leads to abnormal development due to early regulatory programs repeating themselves in later stages of development [Ambros, 1989]. At that time, it was a major surprise that this genetic locus did not encode a protein but gave rise to two small RNA molecules that are partly complementary to multiple sites in the 3'-UTR of *lin-14*, another gene implicated in developmental processes. These RNAs are conserved in multiple nematode species and they were shown to decrease LIN-14 protein levels without affecting mRNA levels [Lee et al., 1993; Wightman et al., 1993].

First, this was not believed to be a widespread mechanism [Bartel, 2004], but years later, *let-7*, another ncRNA implicated in developmental processes of *C. elegans* was found to be conserved throughout the metazoan clade [Pasquinelli et al., 2000]. Up until now, hundreds of these ncRNAs, which were later termed *microRNAs*, were cloned and sequenced [Lagos-Quintana et al., 2001; Landgraf et al., 2007]. Due to the advent of NGS, microRNAs can now be sequenced on large-scale and NGS is commonly applied to discover new microRNAs and to profile their expression levels in various cell-types and conditions [Berezikov et al., 2006; Morin et al., 2008; Witten et al., 2010]. In particular, the centralized microRNA repository miRBase [Griffiths-Jones, 2004], version 19, lists 25,141 mature microRNAs from 193 species, for instance 370 in *C. elegans*, 422 in the fruit fly *Drosophila melanogaster*, 1,281 in mouse and 2,042 in human. Almost at the same time, a cellular mechanism called RNA interference (RNAi) was discovered in *C. elegans* [Fire et al., 1998]. The uptake of exogenous double-stranded RNA molecules leads

to a knock-down of complementary mRNAs. This gene silencing is mediated by another class of ncRNAs called siRNAs. The discovery of RNAi led to the development of potent research tools also for mammalian cell lines [Elbashir et al., 2001] and was awarded a Nobel prize in 2006. Later, it became clear that siRNAs and microRNAs share large parts of their maturation pathway as well as their effector complex called RISC [Bartel, 2004; Kim et al., 2009b].

1.2.2 MicroRNA biogenesis

Canonical mammalian microRNAs are transcribed by RNA polymerase II, often as clusters of many microRNAs on the same primary transcript. After transcription, short hairpin structures called pre-microRNAs of ~ 65 nt length are cleaved out of the primary transcript by the RNase III Drosha, which are subsequently transported out of the nucleus by the nuclear export factor exportin 5. In a second processing step, the hairpin loop is cleaved from the pre-microRNA by another RNase III, Dicer, which is in complex with Argonaute (AGO). Finally, one strand of the remaining duplex, the mature microRNA, resides on AGO1-4 and the RNA induced silencing complex (RISC) is assembled, whereas the other strand is rapidly degraded (reviewed in Bartel [2004]; Kim et al. [2009b]). This canonical biogenesis pathway is widely conserved and orthologs of Drosha, Dicer and AGO can be found in mouse, flies and nematodes.

Intriguingly, there are many alternative roads that can be taken by microRNAs: Most prominently, many microRNAs are not transcribed from their own primary transcripts but are located in introns of protein coding mRNAs. The processing of these so-called *mirtrons* may be very similar to canonical microRNAs, i.e. Drosha cleaves before or after splicing is complete and the hairpin enters the canonical pathway. Alternatively, mirtrons may mature without Dicer: The spliceosome may directly produce pre-microRNAs, or additional nucleotides upstream or downstream of the hairpin may be trimmed by 5' or 3' exonucleases [Kim et al., 2009b; Ruby et al., 2007; Ladewig et al., 2012].

Moreover, various other alternatives of the canonical pathway have been identified including microRNAs that circumvent Drosha and/or Dicer processing by utilizing tRNaseZ [Bogerdt et al., 2010], the integrator complex [Cazalla et al., 2011] or AGO [Cheloufi et al., 2010; Yang et al., 2010; Cifuentes et al., 2010] or microRNAs that are derived from other ncRNAs such as tRNAs [Haussecker et al., 2010] or snoRNAs [Ender et al., 2008; Taft et al., 2009].

In addition to microRNAs, several other ncRNAs have been identified and implicated in regulation, most prominently piRNAs and endo-siRNAs that share parts of the microRNA biogenesis pathway or are similar in their regulatory mechanisms [Ghildiyal and Zamore, 2009; Kim et al., 2009b].

1.2.3 Regulatory mechanisms of microRNAs

As indicated above, microRNAs are important contributors to post-transcriptional regulation. The canonical model of microRNA action is that microRNAs recognize binding sites in the 3'-UTR of a target mRNA by the so-called seed (microRNA bases 2-7 or 2-8), and RISC, which is associated with the microRNA, then downregulates protein expression by inhibiting translation or inducing mRNA degradation [Bartel, 2009]. The notion of a microRNA seed

was first postulated by results from high-throughput experiments, where mRNA expression was measured differentially for cells overexpressing a certain microRNA as compared to control cells lacking this microRNA using microarrays [Lim et al., 2005; Grimson et al., 2007]. The importance of the seed was later corroborated by high-resolution three-dimensional crystal structures of AGO microRNA complexes, which showed that the seed bases are solvent exposed in contrast to the other microRNA bases [Schirle and MacRae, 2012].

However, the existence of a seed site within a 3'-UTR, i.e. a sequence that is reverse complementary to a microRNA seed, is neither sufficient nor necessary for target recognition: If every hexamer seed site, i.e. complementary to bases 2-7 of a microRNA, could indeed be targeted by a microRNA, then a microRNA would have target sites about every 4 kilobases (kb) by chance or in about every fourth transcript (assuming a mean 3'-UTR length of 1000 nt). Moreover, since there are about 1,300 distinct human seeds annotated in miRBase version 19 [Griffiths-Jones, 2004], every 3'-UTR would be fully covered by seed sites by expectation. This is not the case and there are several lines of evidence suggesting that additional factors such as target site location [Grimson et al., 2007], additional basepairing at the microRNA 3' end [Brennecke et al., 2005], target site accessibility [Kertesz et al., 2007] or other factors such as RNA binding proteins [Jacobsen et al., 2010] or microRNA and mRNA copy numbers [Ben-Moshe et al., 2012] play important roles in distinguishing functional target sites from non-functional seed sites. In addition, not even the most general seed site (microRNA positions 2-7) is necessary for seed binding: It has been shown that GU wobbles within the seed-seed site duplex may not necessarily destroy efficient regulation [Didiano and Hobert, 2006] and that microRNA target recognition may also be mediated by other parts of the microRNA than the canonical seed and lead to target downregulation [Shin et al., 2010]. Furthermore, this canonical model is also challenged from another direction by recent studies, since not only 3'-UTR sites seem lead to efficient regulation, but also sites located in protein coding sequences [Tay et al., 2008; Duursma et al., 2008; Reczko et al., 2012; Hausser et al., 2013].

The canonical model introduced above dictates that target protein expression is downregulated upon binding of the microRNA. Various mechanistic explanations for this downregulation have been proposed, including endonucleolytic cleavage (slicing) by AGO2, deadenylation and/or decapping followed by rapid mRNA degradation, deadenylation leading to diminished mRNA circularization, inhibition of translation initiation or elongation, co-translational degradation of the growing peptide chain or sequestering the target mRNA to P-bodies (reviewed in Eulalio et al. [2008]; Pasquinelli [2012]). Some of these modes of action have been challenged [Kozak, 2008], and even if unifying models have been proposed [Djuranovic et al., 2011], recent large-scale studies still report varying estimates about the relative importance of the various mechanisms [Guo et al., 2010; Djuranovic et al., 2012; Bazzini et al., 2012; Mishima et al., 2012]. To make things even more complicated, there is increasing evidence that microRNAs are not only able to downregulate their targets but also that they may lead to an upregulation either directly or indirectly (reviewed in Vasudevan [2012]).

Another important mechanistic aspect of microRNA-mediated regulation is the magnitude of the impact on protein levels. Early examples indicated that microRNAs act in a switch-like manner: For instance, the seminal work on the regulation of LIN-14 by lin-4 showed a complete abrogation of LIN-14 protein levels in lin-4 wildtype individuals as compared to lin-4 mutants

[Lee et al., 1993; Wightman et al., 1993]. Also, another nematode microRNA, *lisy-6*, has been shown to be decisive for the cell fate of the receptor neurons ASE left and ASE right [Johnston and Hobert, 2003]. However, later high-throughput experiments indicated that microRNAs generally lead to mild but widespread repression of gene expression [Lim et al., 2005; Grimson et al., 2007; Baek et al., 2008; Selbach et al., 2008]. A recent study revealed that the magnitude of the effect of microRNAs is dose-dependent ranging from mild repression for highly expressed genes to a switch-like behavior for mRNAs expressed below a certain threshold [Mukherji et al., 2011].

1.2.4 Biological function of microRNAs

As indicated above, microRNAs give rise to a whole layer of gene regulation in addition to transcriptional regulation by TFs. From a functional point of view, microRNAs share many aspects with TFs: For instance, in both cases there is a many-to-many relationship between regulator and target genes, i.e. a microRNA as well as a TF usually regulates multiple target genes and each gene may be regulated by multiple microRNAs as well as TFs. In addition, the expression of both microRNAs and TFs is tightly regulated itself, giving rise to intricate regulatory networks [Hobert, 2008].

MicroRNAs have important functions in development. Most prominently, the regulation of LIN-14 by the microRNA *lin-4*, which gave rise to the discovery of microRNAs, is essential for normal larval development in *C. elegans* [Lee et al., 1993; Wightman et al., 1993] and regulation of COG-1 by the microRNA *lisy-6* determines left/right asymmetry of *C. elegans* taste receptor neurons [Johnston and Hobert, 2003]. Further examples of such essential roles of microRNAs in developmental processes can be found in Wienholds and Plasterk [2005]. In addition, high-throughput experiments showed that microRNAs are differentially expressed during differentiation of embryonic stem cells [Morin et al., 2008] and exhibit tissue specificity in general [Landgraf et al., 2007; Sayed and Abdellatif, 2011]. Thus, microRNAs are implicated in the regulatory networks that determine differentiation events and cell fate [Ivey and Srivastava, 2010].

There is also evidence that microRNAs play important roles in evolution, similarly to TFs [Chen and Rajewsky, 2007]. Some microRNAs show extreme patterns of evolutionary conservation, e.g. *let7* is conserved from human to worm corresponding to an age of more than 600 million years [Pasquinelli et al., 2000; Wheeler et al., 2009], and microRNA families have been added to the regulatory repertoire especially of higher organisms without frequent substitutions or secondary loss [Wheeler et al., 2009]. Importantly, tissue specific expression of those microRNA families is maintained over evolution [Christodoulou et al., 2010]. Thus, microRNAs may have essential functions in speciation and may contribute to macro-evolution and the development of tissues and complex body plans. Intriguingly, the diversity of microRNAs in the human brain as compared to our closest evolutionary relatives, chimpanzees, suggests that they may even have contributed to the evolutionary development of higher brain functions found in humans [Berezikov et al., 2006].

Another functional aspect of microRNAs concerns disease: In an overwhelming amount of studies it has been shown that microRNA related alterations are associated with diverse types of

cancer (e.g. reviewed in Farazi et al. [2011]). Interestingly, different cancer types show different and specific expression patterns of microRNAs, and, thus, microRNAs are promising candidates for prognostic and diagnostic markers for cancer [Calin and Croce, 2006; Witten et al., 2010; Farazi et al., 2011].

Furthermore, the initial hypothesis for the main function of RNAi was defence against viral infection [Jeang, 2012]. This has been confirmed in plants, insects and nematodes, but evidence in mammals is still lacking [Kincaid and Sullivan, 2012]. Intriguingly, viruses also exploit the RNAi machinery for pro-viral purposes, either by regulating host microRNAs or encoding their own microRNAs [Kincaid and Sullivan, 2012]. Since the discovery of microRNAs in Epstein-Barr virus (EBV), a human herpes virus, by Pfeffer et al. [2004], microRNAs have been identified in the genomes of various viruses, where the herpes viruses comprise the class with the highest number of known microRNAs. Similarly to the mimicry of host proteins by several viral proteins, there are a few microRNAs that share seed sequences with host microRNAs, but most viral microRNAs do not show homology to any host microRNA [Kincaid and Sullivan, 2012]. Functional understanding of viral microRNA function is still lacking, but the few existing examples indicate functions in preventing apoptosis, immune evasion and regulation of viral genes [Kincaid and Sullivan, 2012].

1.2.5 Bioinformatics for microRNAs

Computational approaches related to microRNA biology can be grouped into two categories [Mendes et al., 2009]: microRNA gene identification and microRNA target identification.

For the identification of microRNA genes, i.e. loci on the genome that give rise to functional mature microRNAs, mainly three criteria have been used, often combined and using machine learning (reviewed in Mendes et al. [2009]). First, since known pre-microRNAs exhibit a characteristic secondary structure, i.e. a hairpin of ~ 65 nt, secondary structure predictions or features derived from the minimal free energy structure such as number and size of internal loops or the free energy have been shown to be potent characteristics for microRNA gene finding. Later, a mechanistic explanation for the necessity of the hairpin structure has been found, since both the microprocessor complex and Dicer have structural requirements for their substrates [Han et al., 2006; MacRae et al., 2007]. Specifically, the microprocessor complex can only cleave hairpin loops of a certain length and internal loop composition and Dicer recognizes specific double stranded structures. Second, many microRNAs are broadly conserved [Pasquinelli et al., 2000; Wheeler et al., 2009] and highly specific conservation patterns across the precursor hairpin have been found [Stark et al., 2007]. Consequently, evolutionary conservation of hairpin structures have been shown to be a potent filter for true microRNA genes [Stark et al., 2007]. And finally, microRNA obviously have to be expressed to be functional and, consequently, NGS has been utilized to identify novel microRNAs [Berezikov et al., 2006; Morin et al., 2008; Witten et al., 2010].

Each of these criteria entails certain shortcomings: As described above, not all microRNAs and regulatory small RNAs in general are processed by Drosha/Dicer and, thus, a hairpin secondary structure may not be necessary. Also, not all microRNAs are conserved and evolutionary late microRNAs are not less interesting than widely conserved ones. Thus, a method to find regulatory

small RNA without relying on secondary structure or conservation may be of great benefit. In addition, only a relatively small fraction (one to two third) of typical deep sequencing data of small RNAs corresponds to microRNA reads [Berezikov et al., 2006; Morin et al., 2008; Witten et al., 2010] and the function of the remaining part is largely unknown and usually ignored in analyses. Therefore, we developed ALPS (see chapter 3 and ref. [Erhard and Zimmer, 2010]), which is a method to classify ncRNAs with respect to similar deep sequencing read patterns. ALPS can be used to identify unannotated regions in the genome that shows similar read patterns as microRNAs or other regulatory RNAs without relying on secondary structure predictions or conservation. Furthermore, it can be used to cluster all expressed ncRNAs according to their read pattern.

The second category for microRNA related computational approaches is the identification of microRNA targets. This can either be done by means of prediction, by analyzing experimental data or combinations thereof. Since the discovery of microRNAs, a plethora of microRNA target prediction methods has been proposed (e.g. reviewed in refs. [Thomas et al., 2010; Sethupathy et al., 2006; Ritchie et al., 2009; Mendes et al., 2009]). In principle, prediction methods first identify a set of possible sites (e.g. by identifying seed sites) and then apply certain filtering criteria such as target site location [Grimson et al., 2007], additional basepairing at the microRNA 3' end [Brennecke et al., 2005], target site accessibility [Kertesz et al., 2007] or microRNA and mRNA copy numbers [Ben-Moshe et al., 2012]. Often, these additional criteria are integrated into the prediction using machine learning techniques [Betel et al., 2010; Sturm et al., 2010]. However, in general predicted targets of microRNAs are not deemed highly reliable [Ritchie et al., 2009; Thomas et al., 2010], indicating that important criteria for finding true microRNA targets are currently still missed.

As a remedy to those unreliable predictions, several experimental high-throughput techniques have been proposed to discover microRNA targets, either based on expression profiling upon microRNA overexpression or knock-down [Lim et al., 2005; Grimson et al., 2007; Baek et al., 2008; Selbach et al., 2008], or based on biochemical isolation of RISC in association with target transcripts (RIP-Chip/RIP-seq, see refs. Mukherjee et al. [2009]; Hendrickson et al. [2008]; Karginov et al. [2007]; Stoecklin et al. [2008]; Landthaler et al. [2008]) or target sites (AGO-CLIP, see refs. Chi et al. [2009]; König et al. [2010]; Hafner et al. [2010]). In order to extract bone-fide target or target sites from such kind of data, computational methods are necessary for proper analysis. In chapters 5 and 4 I describe methods that I developed to analyze RIP-Chip and AGO-CLIP data, respectively [Erhard et al., 2013b,a].

Apart from methods that belong to the two categories introduced by Mendes et al. [2009] (see above), I developed further computational methods that are related to microRNA biology. First, microRNAs and ncRNAs in general have been implicated in the regulation of alternatively spliced transcripts. There are examples of indirect regulation of alternative splicing either by microRNAs targeting splicing factors such as *nPTB* by miR-133 during muscle development [Boutz et al., 2007], PTBP1 by miR-124 during brain development [Makeyev et al., 2007] or CELF proteins by miR-23 during heart development [Kalsotra et al., 2010] or by target sites located on alternative exons [Tay et al., 2008; Duursma et al., 2008; Yang et al., 2012]. Additionally small RNAs may even be directly contributing to regulation of splicing, e.g. by blocking splice sites, the branch point or other regulatory elements important for differential

splicing [Khanna and Stamm, 2010]. Thus, a systems biology approach to analyze differential spliced genes in large-scale is to examine appropriate high-throughput data for indications of exons that behave differently than other exons from the same gene in terms of quantification fold changes. There are several methods already available for RNA-seq data [Richard et al., 2010; Trapnell et al., 2013], however, ultimately, differential splicing only matters on protein level. Thus, I investigated, to which extent high-throughput SILAC based mass spectrometry can be used to infer differential splicing [Erhard and Zimmer, 2012], which is described in chapter 7.

Second, an important aspect of transcriptional regulation is context-dependence: One of the most striking results from the ENCODE project [Consortium, 2012b] is that the binding of transcription factors (TFs) does not only depend on the presence of the TF, but also on other, context-dependent factors and that this is a general feature of TFs-mediated regulation [Thurman et al., 2012; Neph et al., 2012b]. Context-dependent regulation leads to a complex rewiring of the cells regulatory network dependent on the context [Gerstein et al., 2012]. Whether or not context-dependence is also a general feature of post-transcriptional regulation mediated by microRNAs has not been investigated so far in large-scale. Thus, I integrated several high-throughput datasets measured for the same system and found strong evidence for a widespread context-dependence of microRNA-mediated gene regulation (see chapter 6 and ref. Erhard et al. [2013c]).

Finally, the ultimate goal in systems biology is a predictive understanding of a whole system. This predictive understanding can be achieved if a detailed model of a system can be constructed, which is able to accurately predict its behavior. Such a model is often based on a network or graph and predictions are made by means of simulation. Two components must be established for the simulation of such a network involving microRNA-mediated regulation: a mathematical model that describes the regulatory mechanisms of microRNAs and the network itself including the information which microRNAs target which mRNAs and parameter values for the mathematical model. Since mechanistic aspects of microRNA-mediated regulation is still under heavy debate [Djuranovic et al., 2011; Eulalio et al., 2008; Guo et al., 2010; Kozak, 2008; Mishima et al., 2012], mathematically modeling these mechanisms is still in its infancy although a few attempts have already been made [Khanin and Vinciotti, 2008; Morozova et al., 2012; Eduati et al., 2012]. However, due to various high-throughput methods that are now commonly applied in microRNA research, a multitude of target information and parameter values will be readily available in the near future. Bioinformatic tools will be necessary to cope with all these data and to be able to establish useful models for microRNA-mediated regulation. Thus, already in my Diploma thesis, I started the development of PNMA, a comprehensive modeling platform that is based on a general graph based model, Petri Nets [Murata, 1989], and is highly flexible in the mathematical model that is used for simulation: In my diploma thesis, I integrated Fuzzy logic into PNMA to describe the mathematics for interactions, which is beneficial when only rough knowledge is available about the simulation parameters. However, when more detailed data is available, more detailed simulations are possible. Therefore, I integrated FERN [Erhard et al., 2008] into PNMA, which is a framework for stochastic simulation of mass action kinetics. This is described in chapter 8.

1.3 Herpes viruses

Herpes viruses are double-stranded DNA viruses that are extremely widespread among the human population and other mammals. There are several species that can infect a wide range of host organisms. They have in common a relatively large linear genome of 124-230 kb and their characteristic hallmark is their ability for life-long persistence in a latent form. Usually herpes viruses do not cause life-threatening diseases but some species may contribute to cancerogenesis in immune suppressed patients [Knipe et al., 2007].

Mature herpesvirus virions vary in size from 120 to as much as 260 nm. The virions are composed of the core containing the genome, an icosahedric capsid built of viral proteins, the tegument containing several viral proteins and an envelope with up to 1000 glycoproteins [Knipe et al., 2007]. According to the NCBI taxonomy database (accessed 15.2.2013), 51 genomes from the family *herpesviridae* are sequenced, including all eight known human herpes viruses. Herpes virus genomes contain 70-200 protein coding genes, of which 40 core genes are conserved across the whole family. Most viral genes do not contain any introns, overlapping genes are common within the genomes and many proteins are multifunctional. Their associated functions include DNA replication, packaging of viral genomes, viral replication, immune evasion, establishment of latency [Knipe et al., 2007].

Importantly, all herpes viruses excluding Varicella-Zoster Virus (VZV) [Umbach et al., 2009] encode a set of microRNAs. In contrast to viral proteins, with a few exceptions, viral microRNAs are not conserved across species but are believed to fulfill similar purposes [Kincaid and Sullivan, 2012].

1.3.1 Phylogeny

Based on their protein sequences 702 herpes viruses are classified into the order *herpesvirales* according to the NCBI taxonomy database (accessed 15.2.2013, see Figure 1.2). The order comprises three families, *herpesviridae* that includes mammalian, avian and reptilian viruses, *alloherpesviridae* including fish and amphibian viruses and *malacoherpesviridae* that consists of a single herpes virus infecting a certain oyster species [Davison et al., 2009].

There is no protein conserved within the order herpesvirales that is not also found in other viruses apart from herpes viruses [Davison et al., 2009], but 40 genes are conserved within the family herpesviridae. In addition, the genomic organization is highly conserved. Thus, it is believed that there was a common ancestor of those herpes viruses that already contained ancestral variants of those 40 genes [Knipe et al., 2007].

The family herpesviridae consists of three subfamilies, the alpha-, beta- and gammaherpesvirinae. Generally, alphaherpesvirinae are neurotropic and have relatively short reproductive cycles, whereas betaherpesvirinae have a broad range of host cells and slow reproduction. The gammaherpesvirinae mainly infect lymphocytes. Each subfamily is further split into several genera, each of which consists of several herpes virus species. Human specific herpes viruses can be found in all three subfamilies: Herpes Simplex virus 1 (HSV1) and Herpes Simplex virus 2 (HSV2) belong to the alphaherpesvirinae together with VZV. The betaherpesvirinae

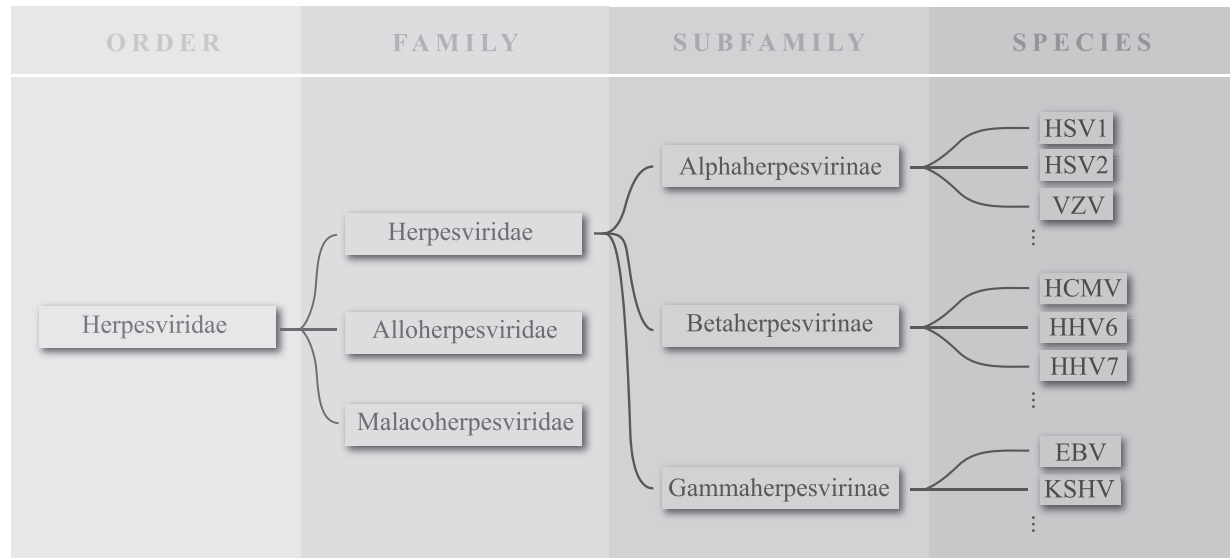


Figure 1.2: Phylogeny of herpes viruses. All herpes viruses are classified into the order herpesvirales. The family herpesviridae contains all mammalian herpes viruses, alloherpesviridae include fish and amphibian viruses (omitted here) and malacoherpesviridae consist of a single virus infecting oysters. The subfamilies are further subdivided into genera which are also omitted here. Only the human herpes viruses are included.

include the Human Cytomegalovirus (HCMV) and Human Herpes viruses 6 and 7. The human gammaherpesvirinae are EBV and Kaposi's Sarcoma-associated Herpesvirus (KSHV). Many of those have close relatives that infect other mammalian species, for instance Murine Cytomegalovirus (MCMV), which is closely related to HCMV, infects mice and also belongs to the betaherpesvirinae. Between HCMV and MCMV, 70 proteins are conserved [Knipe et al., 2007].

1.3.2 Life cycle, symptoms and prevalence

As indicated above, herpes viruses are characterized by two modes of infection: In latency, the viral genome remains in circular form in the nucleus of the host cell, where only a small subset of genes including microRNAs is expressed. It is able to reactivate, often upon cellular stress and to transition into a lytic phase, where the virus replicates DNA and creates its structural proteins. Mature virions are then assembled in the cytoplasm until the host cell bursts, thereby releasing the viral progeny [Knipe et al., 2007].

Thus, herpes viruses are able to cause life-long infections in general. Some species cause certain clinical symptoms upon primary infection, e.g. VZV causes chickenpox upon infection during childhood and often shingles upon infection of adults. Other herpes viruses cause frequently recurring symptoms such as cold sores by HSV1. Infection by other herpes virus species may also stay completely unnoticed, e.g. typically for HCMV. In immunosuppressed patients, however, herpes viruses are associated with cancer.

Due to their effective strategy of latent infections, herpes viruses have relatively high prevalence in the human population in general. For instance, 80% of the human population is infected with HSV1 and more than 90% are infected with EBV, which is relatively uniform across demographic parameters such as sex, ethnicity or prosperity. KSHV, which is often occurring in AIDS patients and not only associated with its namesake Kaposi's sarcoma but also with primary effusion lymphoma and Castleman's disease, occurs in less than 5% in Northern America and Europe but has a prevalence of more than 90% in Africa.

1.3.3 Viral microRNAs

An intriguing feature of most herpes viruses is that they encode not only protein coding genes on their genome but also microRNAs [Pfeffer et al., 2004; Kincaid and Sullivan, 2012]. In general, they are highly conserved across strains from the same viral species but interestingly not between different species [Cullen, 2010; Kincaid and Sullivan, 2012]. There are notable exceptions for closely related herpes viruses, e.g. 22 of the 25 EBV microRNAs are conserved in the related Rhesus monkey Lymphocryptovirus (rLCV), but none of the known human specific herpes virus microRNAs is conserved in any other human specific virus. Importantly, this is in sharp contrast to several widely conserved herpes viral core proteins. Interestingly, even though microRNAs are not conserved by their sequence, their genomic localization is similar among most herpes viruses, which indicates that a common ancestor may already have encoded microRNAs. Thus, microRNAs may be a causative agent for viral speciation [Kincaid and Sullivan, 2012].

Most of the human herpes viruses encode dozens of microRNAs (see Table 1.1), with the exception of VZV [Umbach et al., 2009]. An explanation for the lack of microRNAs in VZV is still missing. It is possible, although unlikely, that the expression levels of existing VZV microRNAs is below the detection limit of the sequencing experiments of Umbach et al. [2009], but it could also be that VZV microRNAs are only expressed during lytic infection, which is difficult to establish in cell culture [Umbach et al., 2009].

Functions of viral microRNAs are largely unknown, but they participate in immune evasion, avoidance of apoptosis and maintenance of latency [Cullen, 2006, 2010; Kincaid and Sullivan, 2012]. For instance, even if respective microRNAs are not homologous, three different human herpes viruses (HCMV, EBV and KSHV) have been shown to target the pro-apoptotic host gene BclAF1 [Kincaid and Sullivan, 2012]. Utilizing microRNAs instead of TFs for regulation may be highly beneficial for the virus: In contrast to proteins, microRNAs are invisible to the immune surveillance system [Cullen, 2006, 2010].

In addition to microRNAs encoded on their own genome, herpes viruses may also exploit host microRNAs for their purposes. For instance, in EBV infected cell lines, the human microRNA hsa-miR-155, which is implicated in cellular processes such as apoptosis and proliferation, is highly induced [Speck and Ganem, 2010; Cullen, 2010]. Intriguingly, while the related gamma herpes virus KSHV does not induce hsa-miR-155 expression upon infection, its kshv-miR-K12-11 is a so-called seed homologue of this host microRNA, i.e. it shares the same seed sequence and should thus recognize similar targets.

Due to their only relatively recent discovery, there are many open questions regarding herpes-viral microRNAs, as indicated above. These include, but are not limited to, which microRNAs

Table 1.1: Experimentally discovered microRNAs of human herpes viruses (according to miRBase [Griffiths-Jones, 2004] version 19).

Virus	Family	pre-microRNAs	mature microRNAs
HSV1	α	17	26
HSV2	α	18	24
VZV	α	0	0
HCMV	β	11	17
HH6	β	4	8
HH7 ¹	β	-	-
KSHV ²	γ	13	25
EBV	γ	25	44

¹ No experimental data is available for Human herpesvirus 7, in contrast to VZV, where several cell lines were considered but no microRNA was found

² Often, only 12 KSHV pre-microRNAs are accounted for, since kshv-miR-K12-10a and kshv-miR-K12-10b are highly similar.

are encoded by herpes viruses and when are they expressed; what are the targets of those microRNAs and of dysregulated host microRNAs; and what is the biological function of these microRNA/target interactions. The purpose of our project *Pathogenic role of miRNAs in herpesvirus infection*, which has been funded by the German Bundesministerium für Bildung und Forschung in the context of the NGFN-plus programme, is to approach these questions with a combination of high-throughput methods, bioinformatics and wet-lab validation. In the context of this project, diverse high-throughput datasets have been generated by our collaboration partners, and the herein presented method have been developed to answer specific questions using these high-throughput datasets. In the next chapter, I will give a short overview about the experiments that have been performed and indicate how the methods were applied to answer specific questions.

Chapter 2

Datasets

2.1 Primer on experimental techniques

Omics experiment rely on high-throughput technologies to produce massive amounts of data that are relevant for the biological system under consideration. The most widely used high-throughput technologies, which were also used in the context of our NGFN project, are DNA microarrays, NGS and mass spectrometry. Depending on the nature of the samples that are subjected to one of those technologies, various kinds of data can be obtained. For instance, NGS has heavily contributed to the success of the ENCODE project [Consortium, 2012b] and allowed to produce genome-wide data for multiple cell lines of DNase hypersensitive sites [Thurman et al., 2012], of footprints of DNA binding proteins [Neph et al., 2012b], of transcription start sites and full length transcripts of protein coding genes and ncRNAs [Djebali et al., 2012], of chromatin modification sites [Arvey et al., 2012], binding sites of specific TFs [Landt et al., 2012] and many more [Consortium, 2012b].

2.1.1 Microarrays, metabolic labeling and RIP-Chip

Before NGS became available, DNA microarrays (also called DNA chips) were the method-of-choice, when the identity and quantity of DNA or RNA molecules had to be determined. In principle, a DNA microarray consists of a huge amount of DNA probes immobilized on a solid surface, where each of the probes is designed to be reverse complementary to a specific known DNA sequence. For instance, the GeneChip Human Exon ST Array from Affymetrix contain more than 5.5 Million probes directed against more than 1 Million exons or putative exons. In the basic experimental protocol, first DNA or RNA is isolated from the sample, followed by cDNA synthesis (only for RNA). The DNA is labeled using fluorescence dyes and then hybridized to the microarray. After washing, the microarray is scanned by a laser, in essence producing a large table containing fluorescence intensity information for each probe as the data output [Miller and Tang, 2009]. The probe determines the identity of the DNA or RNA fragment, whereas its fluorescence signal intensity corresponds to the fragment's abundance in the sample. Introduced by Schena et al. [1995], the number of publications that are based on microarray data exploded

in the years after and is further growing [Miller and Tang, 2009]. It is widely applied to profile expression levels of mRNAs in various conditions [Eisen et al., 1998] and for SNP or CNV detection in individual genomes [Gresham et al., 2008; Miller and Tang, 2009].

By using several modifications or different sample preparation techniques, microarrays can also be used to measure additional kinds of data. For instance, by metabolically labeling newly transcribed RNA by the uridine analogue 4sU, it is possible to biochemically separate newly transcribed from preexisting RNA [Dölken et al., 2008]. These RNA fractions, as well as total RNA can be measured using microarrays, which allows to simultaneously study RNA synthesis and decay. The absolute RNA half-life can be computed either by considering the ratio of preexisting to total RNA or the ratio of newly transcribed to total RNA [Dölken et al., 2008] by

$$t_{1/2} = -t \ln 2 / \ln \frac{A_t}{A_0} \quad (2.1)$$

$$= -t \ln 2 / \left(1 - \ln \frac{A_t^*}{A_0} \right) \quad (2.2)$$

Here, A_0 is the measured amount of total RNA, A_t and A_t^* the amount of preexisting and newly transcribed RNA after labeling for time t . The RNA half-life is closely related to the RNA decay rate $\lambda = \ln 2 / t_{1/2}$ and, thus, an extremely interesting parameter in microRNA related research. Consequently, I considered microarray measurements obtained after metabolic labeling to investigate whether context-dependent microRNA/target interactions have context-dependent influence on RNA decay rates (see Erhard et al. [2013c] and chapter 6).

Another application of microarrays is RIP-Chip [Mukherjee et al., 2009; Hendrickson et al., 2008; Karginov et al., 2007; Stoecklin et al., 2008; Landthaler et al., 2008; Dölken et al., 2010]. Here, in the sample preparation, RNA binding proteins (RBPs) are immunoprecipitated using specific antibodies and co-immunoprecipitated RNA is purified. As a control, this is repeated using an unspecific antibody or total RNA is used. Then, both fractions are measured on microarrays. For each gene, an enrichment value can then be computed:

$$e = \frac{A_s}{A_c} \quad (2.3)$$

A_s is the amount of RNA in the specific IP fraction, whereas A_c is the amount of control RNA. In general, binding partners of the RBP have high enrichment values and mRNAs that is not bound by the RBP should have low enrichment values. Furthermore, these enrichment values are quantitative: Higher values indicate, that an mRNA is a stronger target of the RBP, i.e. a large fraction of all expressed mRNAs is bound by the RBP. However, it is not straight-forward to decide on a cutoff on these enrichment values to define a set of reliable targets. Furthermore, the IP in the sample preparation does not always work with the same efficiency, which introduces bias not existing in standard microarray experiments. Both, the decision of a meaningful cutoff and proper normalization accounting for this bias is described in Erhard et al. [2013b] and is topic of chapter 5.

Importantly, the effector complex of microRNA-mediated regulation, a constituent part of the RNA induced silencing complex (RISC) is an AGO protein, for which antibodies are available. Thus, RIP-Chip can also be used to determine microRNA targets. As a consequence, I also considered RIP-Chip measurements of RISC to investigate context-dependent microRNA/target interactions (see Erhard et al. [2013c] and chapter 6).

2.1.2 Sequencing, sRNA-seq and PAR-CLIP

In recent years, NGS has started to outstrip microarrays. The major disadvantage of microarrays is that the sequence of all RNAs or DNAs to be interrogated must be known beforehand, and the millions of probes on a current microarray are still far away from fully covering a complex mammalian genome and transcriptome. Sequencing in general is the process of determining the sequence of DNA (or RNA). The first generation of sequencers was based on the technology developed by Sanger and Coulson [1975] and was for instance used to determine the full sequence of several genomes including the human genome [Lander et al., 2001; Venter et al., 2001]. Several technologies are counted among the *next generation* of sequencers that emerged in the last decade (e.g. review by Fuller et al. [2009]; Mardis [2008]; Ozsolak and Milos [2011]). In comparison to microarrays, in NGS experiments, the identity of DNA fragments is not determined by their hybridization to complementary probes but by directly determining their sequences and the abundance is not quantified by the fluorescence intensity of the labeling dye but by counting the number of observed sequences. Thus, the typical data output of an NGS experiment is a huge file containing millions of sequences, often accompanied by per-base quality scores. For instance, the four libraries of our PAR-CLIP experiment (see below) were sequenced using a single lane on a Illumina Genome Analyzer IIx and yielded 120 million sequencing reads, each 50 base pairs long.

One specific omics experiment that is based on NGS is short RNA profiling [Pritchard et al., 2012]. Here, short RNAs, i.e. of length about 20-24 nt are specifically selected from cell lysates using gel purification and subjected to NGS. Such datasets can be used for the discovery of novel microRNAs [Berezikov et al., 2006; Morin et al., 2008; Friedlander et al., 2008] and to profile the expression levels of all known microRNAs [Calin and Croce, 2006; Witten et al., 2010; Farazi et al., 2011]. As we and others have noticed, however, there are not only microRNAs that are sequenced, but also many other ncRNAs [Erhard and Zimmer, 2010; Langenberger et al., 2012]. In particular, I developed a method that allows to compare and classify ncRNAs with respect to their pattern of sequencing reads from the NGS experiment, which is described in chapter 3.

NGS is not only about to replace microarrays in standard experiments such as expression profiling of mRNAs or short RNAs, but also the above mentioned microarray based assays have been adapted to use NGS instead, for instance to determine mRNA half lives [Windhager et al., 2012; Rabani et al., 2011] or binding partners of RBPs [Zhao et al., 2010]. However, NGS has capabilities that exceed those of microarrays. In particular, by using sequencing, it is not only possible to identify target genes of RBPs, but also to determine the specific target sites with basepair resolution. This is done by so-called crosslinking and immunoprecipitation (CLIP) experiments [Chi et al., 2009; König et al., 2010; Hafner et al., 2010]: Irradiation of cells using UV light with a specific wavelength crosslinks proteins to RNA, i.e. covalent bonds between

RNA bases and amino acids are formed. Then, the RNA is digested enzymatically such that a small footprint of a few dozen nt remains crosslinked to all RBPs. The RBP of interest is then isolated using IP and the protein is digested. Thus, footprints of the RBP remain, which correspond to binding sites and are sequenced using NGS. Importantly, due to these crosslinks, the efficiency of co-immunopurifying target sites is much higher than without crosslinking in RIP-Chip. Unfortunately, there is also a counteracting aspect in most CLIP protocols, since the cDNA synthesis, which is necessary before NGS, is hindered by amino acid residues still crosslinked to the RNA. One of the specific CLIP protocols, iCLIP [König et al., 2010] has a very promising solution for this issue: Only a single reverse transcription primer is used and cDNA synthesis stops at the crosslinking site. This end is then ligated to the other end of the primer, i.e. the cDNA is circularized. Then, the circular cDNA is cut within the primer yielding a single-stranded, linear DNA molecule suitable for PCR and sequencing.

Another recent modification of CLIP is Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP), which has been used by various groups to identify microRNA binding sites [Hafner et al., 2010; Gottwein et al., 2011; Lipchina et al., 2011; Kishore et al., 2011; Skalsky et al., 2012]. For PAR-CLIP, cells are labeled using 4sU, which is specifically and efficiently crosslinked at a different wavelength than used in the original CLIP protocol. Specifically means that effectively only incorporated 4sU and not other nucleosides get crosslinked, and efficiently means that cells can be irradiated using less energy than in normal CLIP experiments. Importantly, crosslinked 4sU is not read as uridine during cDNA synthesis but as cytidine. Thus, after aligning sequencing reads to the reference sequence (genome or transcriptome), characteristic T to C mismatches can be observed in true binding sites. These mismatches constitute a powerful feature to identify true binding sites [Corcoran et al., 2011; Erhard et al., 2013a]. In addition to the identification of true binding sites, for AGO-PAR-CLIP experiments there is another important task in the data analysis: Determine the specific microRNAs that binds to each of the identified target sites. For this task, these conversions can also be exploited (in addition to other features of PAR-CLIP data), since conversions occur at specific positions relative to the microRNA seed site [Erhard et al., 2013a]. How these feature can be utilized to accurately identify true binding sites and the correct microRNA binding there is described in chapter 4.

2.1.3 LC-MS/MS and SILAC

Both technologies, DNA microarrays and NGS are able to interrogate RNA or DNA. For other domains of biological molecules, namely proteins and metabolites, a different technology has been established: mass spectrometry [Ong and Mann, 2005; Cox and Mann, 2007]. In particular, to identify and quantify proteins, a widely used platform is liquid chromatography tandem mass spectrometry (LC-MS/MS) based on stable isotope labelling of amino acids in cell culture (SILAC). Here, proteins of a sample are isolated and digested into peptides. These are then injected into a high resolution tandem mass spectrometer using a chromatographic column. In principle, a mass spectrometer is able to measure the masses of many molecules simultaneously

in a quantitative way ¹. Thus, a mass spectrum is a list of masses with associated intensities. *Tandem* mass spectrometers are able to measure two kinds of mass spectra: First, masses and intensities of all peptides in the mass spectrometer are determined, followed by one or multiple secondary measurements. For each of these secondary measurements (called tandem mass spectra), one peptide is selected, fragmented by collision with an inert gas and peptide fragments are measured. Thus, the output of a typical LC-MS/MS experiment consists of thousands of primary mass spectra and tens of thousands of tandem mass spectra [Cox and Mann, 2008; Michalski et al., 2011].

In general, tandem mass spectra are used to identify peptide sequences [Cox and Mann, 2008; Cox et al., 2011]. After the collision induced fragmentation, the most abundant fragments are prefixes and suffixes of the original peptide. Based on mass differences of these fragments, both the identity of amino acids as well as their sequence can be inferred [Cox et al., 2011]. This inference is often done by comparing the measured spectra to theoretical spectra computed from sequences of known peptides. A decoy peptide approach is often used to assess the false discovery rate (FDR) [Elias and Gygi, 2007]: Experimental spectra are not only compared to computed spectra of real peptides but also of decoy peptides, e.g. randomized or reversed peptide sequences. Obviously, all identified decoy peptides are erroneous identifications. Thus, the fraction of identified decoy peptides should correspond to the fraction of erroneous identifications of real peptides. In a typical mass spectrometry experiment 60-80% of all measured tandem mass spectra can be assigned with an FDR of 1% [Cox et al., 2011; Michalski et al., 2011].

The primary mass spectra are used to quantify peptides [Ong and Mann, 2005; Cox and Mann, 2008]. This is often done after metabolic labeling (SILAC): To compare protein levels in two cell cultures, one of them is grown on a medium with heavy isotopes of the amino acids Arginine and Lysine, which are then incorporated into proteins. Then, proteins are isolated from both cell cultures and mixed. This protein mix is digested using Trypsin, which specifically cleaves after Arginine or Lysine. Hence, each tryptic peptide apart from the C-terminal peptide contains either an Arginine or Lysine and thus, the source cell culture of each peptide molecule can be determined by its mass. Since the heavy isotope does not alter the physico-chemical properties of the peptide, both, the light and the heavy peptide elute at the same time from the chromatographic column. Therefore, in the primary mass spectra, peptide pairs are observed that are characterized by a mass shift corresponding to the mass difference of the labeled amino acid. The ratio of the intensities of peptide pairs then corresponds to the peptide fold change between the two cell cultures. Importantly, it is also possible to compare three cell cultures at once using two different heavy isotopes [Cox and Mann, 2008] and systems are available to perform SILAC in vivo [Zanivan et al., 2012].

2.2 Cell lines and available datasets

Studies about human herpes viruses are usually conducted in in-vitro systems using stable cell lines [Knipe et al., 2007]. Often, such cell lines are established after extraction from tumor tissue

¹To be precise, the mass over charge ratio of ions is measured.

Table 2.1: Datasets for the cell lines DG75-eGFP, DG75-10/12 and BCBL1. The numbers denote replicate measurements.

Experiment	Parameters	Publication	Comments
Microarray	mRNA levels	[Dölken et al., 2010]	
Microarray (4sU)	mRNA half lives	[Dölken et al., 2010]	Labeling time 60 min
LC-MS/MS	Protein levels	[Erhard et al., 2013c]	SILAC triple labeling
RIP-Chip	microRNA targets	[Dölken et al., 2010]	Ago2 specific antibody
PAR-CLIP ¹	microRNA target sites	[Erhard et al., 2013a,c]	Ago2 specific antibody

¹ No PAR-CLIP was performed for DG75-10/12

(e.g. described by Ben-Bassat et al. [1977]; Renne et al. [1996]), or are immortalized by infection with EBV (e.g. in Skalsky et al. [2012]). Our NGFN-plus project *Pathogenic role of miRNAs in herpesvirus infection* focussed on a few cell lines that were either infected by a specific herpes virus or not. To allow an integrative approach, several high-throughput experiments have been performed in the same cell lines (see chapter 6).

2.2.1 KSHV related cell lines

The main system of cell lines used in this work consists of three cell lines, DG75-eGFP, DG75-10/12 and BCBL1. DG75 is a relatively old B-cell line that is EBV and KSHV negative and has been extracted from a patient with primary abdominal lymphoma [Ben-Bassat et al., 1977]. This cell line has been transduced with a lentiviral vector either expressing enhanced green fluorescent protein (eGFP) or 10 of the 12 KSHV microRNAs ² that are encoded within an intron of the Kaposin gene [Dölken et al., 2010]. BCBL1 is an EBV negative B-cell line extracted from a patient with body-cavity based lymphoma that is latently infected with KSHV [Renne et al., 1996].

Various high-throughput experiments have been conducted for these three cell lines by our collaboration partners in the NGFN-plus project, namely microarray measurements of total, newly transcribed and preexisting RNA (see above), SILAC based LC-MS/MS, AGO-RIP-Chip and AGO-PAR-CLIP (excluding DG75-10/12; see also table 2.1). I analyzed these datasets using the methods I developed (see chapters 5,4 and 7) or already available methods [Dölken et al., 2008; Cox and Mann, 2008] and integrated them in order to investigate the widespread context-dependence of microRNA-mediated regulation (see chapter 6). For the same analysis, I also integrated publicly available PAR-CLIP data from Gottwein et al. [2011], where other KSHV positive B-cell lines were measured (BC1 and BC3). For the evaluation of the accuracy of PARma, I also considered our PAR-CLIP data set in DG75 and BCBL1 as well as the BC1 and BC3 data from Gottwein et al. [2011].

²The missing microRNAs are kshv-mir-K12-10 and kshv-mir-K12-12

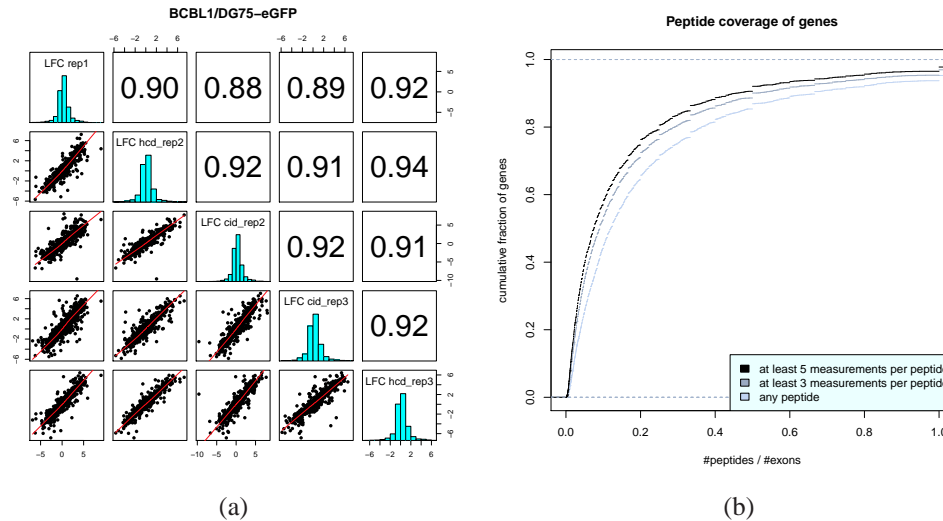


Figure 2.1: Proteomics data for DG75-eGFP, DG75-10/12 and BCBL1. In figure 2.1a, the gene log₂ fold change correlations are shown for all pairs of replicates. The values in the upper triangle represent the Pearson correlation coefficient for the respective pair of replicates. All genes with at least two measured peptides were considered. Three independent biological replicates have been measured (rep1-3), two of which have been repeated with different collision techniques (CID, collision induced dissociation; HCD, higher-energy collisional dissociation) and therefore represent technical replicates. The overall correlation of log fold changes between replicates was about 0.9, which indicates a very good reproducibility of the experiments. Figure 2.1b illustrates the sparseness with respect to peptide coverage of the mass spectrometry data. For each gene, the number of identified peptides was divided by the number of exons. Thus, this cumulative distribution shows how many genes have at least a specific average number of peptides per exon. In particular, for only about 15% of all identified genes, at least half of the exons contain an identified peptide on average and for about 50% of all genes, 10% or less of the exons contain a peptide on average. This average number of peptides per exon becomes even lower when only repeatedly measured peptides are considered, which are important for a reliable identification of differential splicing.

In addition to the context-dependence of microRNA/target interactions, I also investigated the effect of herpes viruses, and their microRNAs in particular, on alternative splicing patterns of host genes. To this end, I considered the LC-MS/MS data measured for the three cell lines (see table 2.1). First, I investigated, to which extent SILAC based mass spectrometry data can be used to identify genes that are differentially spliced on protein level (see chapter 7 and Erhard and Zimmer [2012]). Unfortunately, even if the quality of the data was quite good (see Figure 2.1a) and the experiment yielded a deep coverage of the proteome with respect to the number of proteins (peptides mapping to 5247 Ensembl genes could be identified with an FDR of 1%), no promising candidates for differential splicing could be found in the data. The main reason for that is probably the low peptide coverage per protein (see Figure 2.1b). Of course, in order

to reliably identify differentially spliced genes, peptides must be identified that distinguish the differential isoforms. Thus, due to the sparseness of the measurements interesting candidates of differential splicing may have been missed and therefore, the question about differential splicing in these cell lines cannot be answered by the mass spectrometry experiments alone (see chapter 7 and Erhard and Zimmer [2012]).

2.2.2 EBV related cell lines

At the time I developed REA (Erhard et al. [2013b]; see also chapter 5), which is a method to analyze RIP-Chip data, no matching PAR-CLIP data was available for comparison in BCBL1. However, in Dölken et al. [2010], RIP-Chip was not only performed for DG75-eGFP, DG75-10/12 and BCBL1, but also for three other B-cell lines, BL41, BL41/B95.8 and Jijoye. All of these are derived from Burkitt's lymphoma patients and are either herpes virus negative (BL41), infected by EBV (Jijoye) or infected by the EBV strain B95.8 that lost several of its microRNAs (BL41/B95.8). Importantly, HITS-CLIP data was publicly available for Jijoye from another lab [Riley et al., 2012a], which I used for evaluations of REA.

2.2.3 VZV related cell lines

I developed ALPS originally to analyze sRNA-seq data measured for VZV infected MeWo cells, which is a melanoma derived skin cell line. The purpose of this study was to investigate whether VZV indeed does not possess any microRNAs: In Umbach et al. [2009], primary cells from trigeminal ganglia, i.e. nerve tissue, were extracted from deceased patients that did not show any signs of virus reactivation. Thus, these cells represented only the latent stage of VZV infection. Even if in these cells, no microRNAs could be found, VZV could nevertheless possess its own microRNAs, for instance only expressed during lytic infection. Thus, skin cells were infected with VZV, which represent the natural host cells for lytic infection and exhibit the clinical symptoms of VZV (chickenpox and shingles). Short RNAs from cells infected by the v-Oka strain of VZV, as well as from mock-infected and uninfected cells were then subjected to NGS by our collaboration partners. Unfortunately, by close inspection of the sequencing data, I found out that something must have gone wrong with the experiments: Sequencing reads were too short (see Figure 2.2) for all libraries and almost no reads could be reliably mapped to the VZV genome (data not shown).

In the following chapters, all methods that I introduced above are presented in detail. Each of these chapters has already been published in a peer-reviewed journal or is submitted for publication. As a preface for each chapter, I briefly state the status of the publications as well as my contributions.

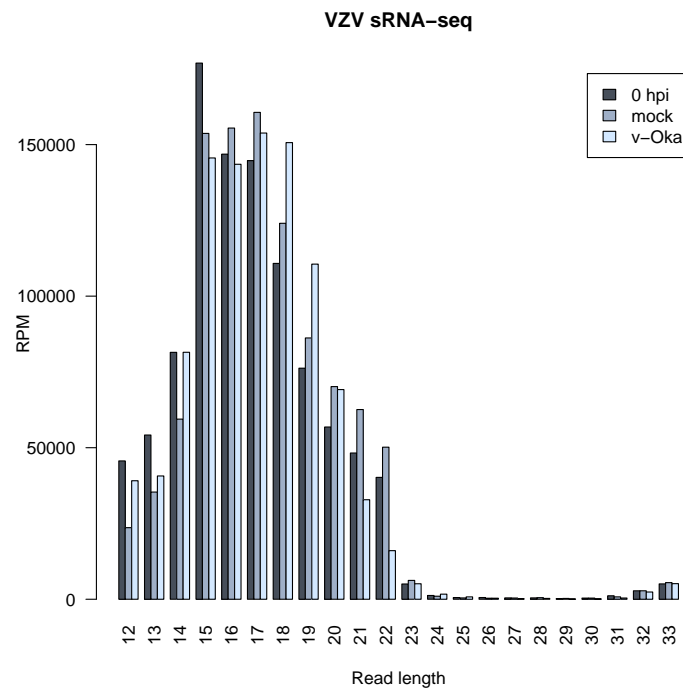


Figure 2.2: Read length distribution of the VZV sRNA-seq experiments. The number of reads per million reads in the library is shown for the three experiments performed (0hpi, uninfected; mock, mock-infected; v-Oka, infected). In all experiments, most read lengths are in the range of 14 to 19, in contrast to usual microRNA sequencing experiments where most reads are typically about 23 bases long [Berezikov et al., 2006; Morin et al., 2008; Friedlander et al., 2008; Witten et al., 2010], indicating severe experimental problems.

Chapter 3

Classification of ncRNAs using position and size information in deep sequencing data

Motivation: As indicated in section 2.2.3, one of the still unresolved questions regarding microRNAs and herpes viruses is whether VZV does or does not encode and express its own microRNAs. All other herpes viruses apart from VZV have been found to express own microRNAs (see above), however, studies specifically investigating possible VZV microRNAs in latently infected cell lines have failed so far to identify any [Umbach et al., 2009]. Thus, our collaboration partners considered a different system, infection of the melanoma derived MeWo cell line. Importantly, other than previous studies, MeWo cell lines represent cell types where VZV exhibits lytic infections and its clinical symptoms (chickenpox and shingles). However, these experiments did not yield any results due to problems with the libraries that were submitted for sequencing (see section 2.2.3). Nevertheless, while analyzing the MeWo sequencing data and, for comparison, data from a published short RNA sequencing study [Morin et al., 2008], I noticed that sequencing reads do not only come from microRNA loci, but also from tRNAs, snoRNA, snRNAs and many other ncRNA classes. Intriguingly, reads seemed to exhibit class specific patterns with respect to relative start positions and lengths. Thus, I systematically investigated to which extent these patterns can be used to distinguish ncRNA classes and to discover putative ncRNAs that exhibit similar patterns to regulatory RNAs such as microRNAs. Notably, this idea has subsequently also been picked up by others [Langenberger et al., 2012].

Publication: This chapter has been published in Bioinformatics [Erhard and Zimmer, 2010] and I presented this work at the European Conference on Computational Biology (ECCB) 2010 in Ghent, Belgium in the proceedings track. Here, I adapted the layout and made minor corrections to the text.

My contribution: I came up with the idea and the method, implemented the method, carried out evaluations and wrote the paper.

Contribution of co-authors: Ralf Zimmer supervised the work and helped to revise the manuscript

3.1 Abstract

3.1.1 Motivation

Small non-coding RNAs (ncRNAs) play important roles in various cellular functions in all clades of life. With next generation sequencing techniques, it has become possible to study ncRNAs in a high-throughput manner and by using specialized algorithms ncRNA classes such as microRNAs can be detected in deep sequencing data. Typically, such methods are targeted to a certain class of ncRNA. Many methods rely on RNA secondary structure prediction, which is not always accurate and not all ncRNA classes have are characterized by a common secondary structure. Unbiased classification methods for ncRNAs could be important to improve accuracy and to detect new ncRNA classes in sequencing data.

3.1.2 Results

Here, we present a scoring system called ALPS (alignment of pattern matrices score) that uses primary information from a deep sequencing experiment, i.e. the relative positions and lengths of reads, to classify ncRNAs. ALPS makes no further assumptions e.g. about common structural properties in the ncRNA class and is nevertheless able to identify ncRNA classes with high accuracy. Since ALPS is not designed to recognize a certain class of ncRNA, it can be used to detect novel ncRNA classes, as long as these unknown ncRNAs have a characteristic pattern of deep sequencing read lengths and positions. We evaluate our scoring system on publicly available deep sequencing data and show that it is able to classify known ncRNAs with high sensitivity and specificity.

3.1.3 Availability

Calculated pattern matrices of the datasets hESC and EB are available at the project website <http://www.bio.ifi.lmu.de/ALPS>. An implementation of the described method is available upon request from the authors.

3.2 Introduction

Next generation sequencing platforms such as Solexa/Illumina, Abi Solid or 454/Roche are extensively used to sequence small RNAs of roughly 14-36 nt length at astonishing rates in various organisms [Morin et al., 2008; Babiarz et al., 2008; Czech et al., 2008; Rathjen et al., 2009; Kato et al., 2009]. For instance, they are used to determine expression profiles of microRNAs, 20-24 nt long RNA molecules, that have emerged in recent years as important post-transcriptional regulators in all known multicellular organisms and that are known to play roles in development, tumorigenesis and viral infection [Bartel, 2004]. Besides microRNAs other small non-coding RNA (ncRNA) classes such as piRNAs [Aravin et al., 2001], snoRNAs [Bachellerie et al., 2002] or scaRNAs [Gerard et al., 2010] have been investigated. Only recently,

454 sequencing revealed the existence of 16 nt long RNA (therefore termed unusual small RNA or usRNAs) in cells infected with KSHV [Li et al., 2009]. usRNAs are derived from both virus and host cell and are associated with the RNA induced silencing complex (RISC). Advances in throughput, accuracy and the ability to sequence longer reads will not only lead to more and more precise detection of already known ncRNA classes, but also to the discovery of new types. It is therefore of great interest to develop methods for automatic classification of ncRNA using deep sequencing data.

Most ncRNAs have very specific structural properties that have been used to classify them [Will et al., 2007], e.g. tRNAs possess a cloverleaf structure, whereas microRNA precursors form stable hairpins. However, these methods rely on the prediction of RNA secondary structure and even for short molecules the current RNA secondary structure energy model is not always able to predict the native structure [Dowell and Eddy, 2004; Doshi et al., 2004]. For instance, the predicted optimal secondary structure of 43 out of 579 murine microRNA precursors in miRbase [Griffiths-Jones et al., 2008] is not an unbranched hairpin (data not shown). This can be explained by the fact that the minimal free energy structure is not necessarily the native one due to unknown modifications or kinetic effects [Higgs and Morgan, 1995]. In the case of de-novo prediction of e.g. microRNAs, the exact pre-microRNA sequence is not known a-priori. Even if the hairpin can be predicted for the pre-microRNA sequence, it could be disrupted, if a few bases upstream or downstream are appended or removed from this sequence. Therefore, multiple windows around a putative microRNA are folded or a local folding tool such as RNALfold [Hofacker et al., 2004] is used. This however necessarily leads to an increased false positive rate since many genomic sequences that do not encode microRNAs can fold into stable hairpins [Bentwich, 2005].

In addition, secondary structure prediction is very sensitive to the exact range that is used for prediction. A microRNA hairpin that can correctly be predicted if one uses the correct genomic range, could be disrupted if a few bases upstream or downstream of the microRNA precursor are included for prediction. Furthermore, according to predictions, many genomic sequences can fold into stable hairpins [Bentwich, 2005] which makes methods using structural features alone quite unspecific.

Deep sequencing offers additional criteria to distinguish ncRNA classes. A typical experimental setup is to determine the content of small ncRNA in a cell under certain conditions. Therefore, only intervals on the genome are considered, where enough sequencing reads have been aligned to. The specific number of reads depends on the tradeoff between sensitivity and specificity. If the experiment aims to identify a special class of ncRNAs, specialized algorithms can be applied that detect specific features of that ncRNA class based on biological knowledge. E.g. in microRNA biogenesis, one strand of the precursor is preferentially included into RISC (the mature microRNA) and the other is rapidly degraded (microRNA star). Considering this bias together with structural microRNA properties can dramatically increase specificity of microRNA detection, as shown before [Morin et al., 2008; Friedlander et al., 2008].

microRNAs recognize their targets by their seed region [Grimson et al., 2007] and, due to their biogenesis, have specific lengths [MacRae et al., 2007]. Both features of microRNAs should be detectable in an excess of deep sequencing reads that align to a specific genomic position and have a specific length. However, this is not always the case for microRNAs in large scale experiments (e.g. [Morin et al., 2008]). The read start position of many microRNAs follow a

32.3. Classification of ncRNAs using position and size information in deep sequencing data

narrow distribution that is often skewed towards the microRNA 3' end and read lengths are often variable (see also Figure 3.1b). Such alternative mature microRNA forms are often referred to as isomiRs [Morin et al., 2008].

In addition to positioning and lengths of reads, distances of reads aligned in close proximity of other reads also carry information about ncRNA classes: At least for animals, the microRNA star should be detectable at a distance of roughly 40 nt to the mature microRNA [Friedlander et al., 2008]. Distance information also helps to distinguish microRNAs from degradation products of other abundant RNA species such as tRNAs or snRNAs (see Figure 3.1). And, most importantly, using this information can help to classify novel ncRNA or ncRNAs that do not possess a characteristic secondary structure.

In this paper, we show how to exploit position and length dependent read patterns to classify ncRNAs. We make no further assumptions about structural and other class specific properties and only consider primary information from the alignment of deep sequencing reads on the genome. Our method ALPS allows to detect microRNAs and other known ncRNA classes with high accuracy and due to its unbiased nature, it also provides a straight-forward way to discover and classify novel ncRNAs. Our approach is complementary to existing methods that rely on structural properties and we expect that their combination with our approach allows to increase their sensitivity and specificity.

3.3 Approach

The starting point for ALPS is the output of a short read aligner (e.g. Bowtie [Langmead et al., 2009] or BWA [Li et al., 2009]) consisting of the positions in the genome, where deep sequencing reads have been aligned to. Then, intervals are identified by clustering these positions such that (1) each interval contains at least m reads, (2) there is no consecutive part of length $> t$ within an interval, that is not covered by a read and (3) t nucleotides downstream and upstream are not covered by a read. The classification problem of ncRNAs using deep sequencing data then is to assign a class label, e.g. *microRNA*, *tRNA*, *snoRNA*, ... to each of these intervals. For a well-annotated organism such as human, mouse or yeast such class labels are already available for many of these intervals in public databases. Then, class labels for the intervals without annotation can be predicted based on similarity to intervals with known annotation, which is often called (semi-) supervised learning. If no or only very few annotations are available for the organism in question, intervals can still be clustered in an unsupervised manner. Both approaches need a way to calculate the similarity between two intervals.

ALPS is such a similarity score computed by an alignment of their so called pattern matrices. These contain the information about the positions and lengths of aligned reads. Since we cannot assume, that all exact distances between aligned reads are always representative for an ncRNA class, we allow gaps in ALPS. For instance, to respect the distance of the mature microRNA and their corresponding microRNA star, our algorithm must be allowed to align the start positions of the two mature microRNAs as well as the start positions of the two microRNA stars, even if the loops of the two precursor microRNAs have different lengths (see also Figure 3.2).

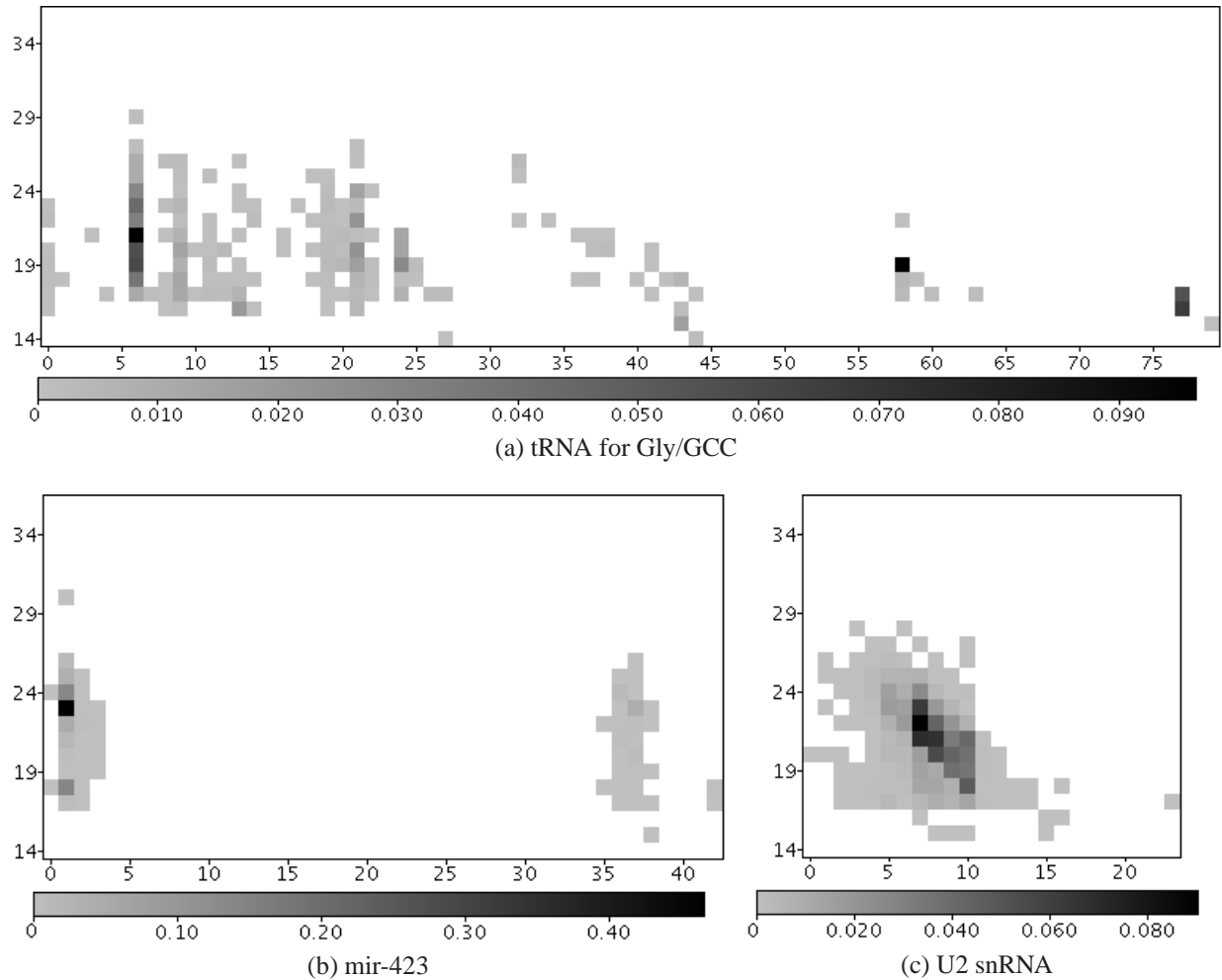


Figure 3.1: Typical position and length dependent pattern matrices for a tRNA, a microRNA and a snRNA. Frequencies of reads starting at position (x axis) and of length (y axis) are visualized in different shades of gray. Note that both the snRNA and the tRNA could easily be mistaken for a microRNA, if only the most abundant read is considered. Graphical representations for all pattern matrices are available on the project website.

Usually, for many intervals, annotations are already available in public databases and these can be used to classify unknown - so far not annotated - intervals similar to them. Generally, ALPS similarities are not biased towards a special class of ncRNAs, since they are only based on the primary data from the deep sequencing experiment. Therefore, used as a distance measure for any unsupervised clustering, the similarity of pattern matrices (ALP score) will find groups of ncRNAs, that exhibit similar distributions (with respect to relative position and length) of deep sequencing reads. If such a distribution is characteristic for an unknown class of ncRNAs, the clustering based on our score should be able to detect it.

In this paper, the focus is not on the detection of unknown classes and hierarchies of ncRNAs but on the detection of already known ncRNA classes to demonstrate the usefulness of our scoring

34.3. Classification of ncRNAs using position and size information in deep sequencing data

system. Based on annotations retrieved from mirBase [Griffiths-Jones et al., 2008], gRNAdb [Chan and Lowe, 2009], Ensembl and Refseq, we identify intervals of known ncRNA classes in published deep sequencing data and benchmark our scoring system based on its ability to reassign an interval to its correct class, after its class label has been removed.

3.4 Methods

To identify the set of intervals \mathcal{I} and their corresponding pattern matrices, we iterate over the sorted read alignments and add a read $r = (r_1, r_2)$ to the current interval $I = (i_1, i_2)$ as long as $i_2 > r_1 - t$, where r_1 and r_2 are genomic start and end of r , respectively, and t is a user-defined tolerance (we use $t = 50$ throughout the paper). Since we do these iterations per chromosome and per strand, each interval spans reads that mapped to one strand of a single chromosome in close proximity to each other and reads of two different intervals are either on different strands or chromosomes or more than t nt apart from each other. An entry $N^I[l, i]$ of the *pattern matrix* N^I of interval I is the number of reads of length l starting at position i in this interval. Positions are according to the strand direction, i.e. if i_1 and i_2 are genomic start and end of an interval on the -strand and a read $r = (r_1, r_2)$ falls into that interval, it contributes to the entry $N^I[r_2 - r_1, i_2 - r_2]$ of the pattern matrix. Since we want to compare pattern matrices for similarity regarding bias of read start positions and lengths frequencies and we have to respect that two ncRNAs of the same class can be expressed at different levels, we normalize each pattern matrix:

$$\tilde{N}^I[l, i] = \frac{N^I[l, i]}{\sum_{l', i'} N^I[l', i']} \quad (3.1)$$

To quantify the similarity of two intervals $I, J \in \mathcal{I}$, we consider their normalized pattern matrices \tilde{N}^I and \tilde{N}^J as sequences of column vectors $(\tilde{N}^I[\bullet, i])_{i=1..|I|}$ and $(\tilde{N}^J[\bullet, j])_{j=1..|J|}$ and compute their optimal alignment. Here we adopt the notation, that $A[\bullet, i]$ is the i th column vector of matrix A . Thus, a column vector is the length distribution of deep sequencing reads, that start at a certain position within the interval. Note, that this distribution is normalized to the proportion of reads that start at this position. The similarity score $S^{I,J}(i, j)$ for aligning position i in interval I to position j in interval J is computed according to

$$S^{I,J}(i, j) = (\tilde{N}^I[\bullet, i])^T \otimes M \otimes \tilde{N}^J[\bullet, j] \quad (3.2)$$

where M is a $L \times L$ matrix (L is the maximal read length). In the simplest case, the identity matrix $M = id_L$ is used and \otimes is the usual matrix multiplication. Then the similarity score is basically just the scalar product of the corresponding column vectors. However, since ncRNA classes are usually not defined by a specific length but by a narrow distribution of lengths, it is reasonable to reward not only exact length matches but also small differences and to penalize large deviations of peaks in the length distributions. Therefore, we use a matrix $M = H^{k,\lambda}$ derived from the sigmoidal function:

$$H[i, j]^{k,\lambda} = h^{k,\lambda}(|i - j|) \quad (3.3)$$

$$h^{k,\lambda}(x) = 1 - \frac{2x^k}{\lambda^k + x^k} \quad (3.4)$$

This matrix rewards differences in read lengths, as long as the absolute difference is at most λ and penalizes all deviations of more than λ . The parameter k describes the steepness of rewards and penalties. The standard sum-product matrix multiplication can also be replaced by a sum-min matrix multiplication. If $M = id_L$ is used and the two column vectors are considered as functions, this score can be geometrically interpreted as their common integral. Again, a hill function derived matrix $H^{k,\lambda}$ can be used to respect length distributions (after negative entries in the matrix have been removed). The ALPS similarity, i.e. the optimal alignment score of the two intervals I and J then is:

$$\hat{s}(I, J) = \max_A \left\{ \sum_{(i,j) \in A} S^{I,J}(i, j) + \sum_{n \in G(A)} g(n) \right\} \quad (3.5)$$

$$g(n) = o + e \cdot n \quad (3.6)$$

The maximum in equation 3.5 is over all possible alignments A of the intervals I and J and $G(A)$ is the set of all gaps in alignment A . Note that the affine gap cost function 3.6 penalizes many short gaps more than few long gaps, which is important for our similarity scoring. We can calculate $\hat{s}(I, J)$ efficiently using the algorithm of [Gotoh, 1982] in time $\mathcal{O}(|I| \cdot |J| \cdot L)$ after a preprocessing of the scoring function S in time $\mathcal{O}(|J| \cdot L^2)$. The preprocessing involves the computations of the second matrix multiplication $M \otimes \tilde{N}^J[\bullet, j]$ for all $j \in [1; |J|]$.

The score in equation 3.5 corresponds to an optimal global alignment. However, we can also define other variants of ALPS similarity: The optimal freeshift (also often called semi-global) alignment score $\hat{s}^f(I, J)$ is given as in equation 3.5 by replacing $G(A)$ by $G^f(A)$, that contains all gaps from $G(A)$ but the longer of the two leading gaps and the longer of the two trailing gaps. Similarly, for the optimal local alignment score, $\hat{s}^l(I, J)$, $G^l(A)$ is used instead of $G(A)$, that contains all but both leading and both trailing gaps. This is equivalent to the usual definition of local alignment, i.e. the optimal global alignment of two subsequences. Note, that we can compute the optimal local and freeshift alignments efficiently using a modified version of the Gotoh algorithm, as suggested in [Smith and Waterman, 1981].

Thus, a scoring system for pairwise ALPS similarities can be described by the 5-tuple $\mathcal{S} = (M, \otimes, o, e, mode)$, where M is the matrix and \otimes the operator for the calculation of the column vector similarity, respectively, o, e are the gap open and gap extend parameters for the affine gap cost function and $mode$ is the alignment mode (global, local or freeshift).

We compute the pairwise ALPS similarities $\hat{s}(I, J)$ for all intervals $I, J \in \mathcal{I}^{m,t}$ that contain at least m reads with tolerance t given a scoring system \mathcal{S} . Then we assign a class to each of the intervals by using annotations from mirBase [Griffiths-Jones et al., 2008], gtRNAdb [Chan and Lowe, 2009], Ensembl and RefSeq. For intervals with multiple assigned annotations, we prioritize annotations according to table 3.1 and we combine similar annotations. All intervals annotated with A are thus partitioned into a cluster C^A . We define the inner and outer similarity scores of class A as the sets

$$D^{inner}(A) = \{\hat{s}(I, J) | I, J \in C^A\} \quad (3.7)$$

$$D^{outer}(A) = \{\hat{s}(I, J) | I \in C^A, J \notin C^A\} \quad (3.8)$$

Using their respective distributions $P^{inner}(A)$ and $P^{outer}(A)$, we can estimate the ability of \mathcal{S} to separate A from all other classes. This means, by using general optimization techniques such

36 3. Classification of ncRNAs using position and size information in deep sequencing data

Table 3.1: Annotations from mirBase, gtRNAdb, Ensembl and RefSeq, ordered by their priority used for the initial class assignment. Similar annotations are combined and the number of respective intervals in the two datasets used for benchmarking is given.

Origin	Annotation	Combined	hESC	EB
mirBase/Ensembl	microRNA	microRNA	103	101
Ensembl	microRNA_pseudogene	microRNA		
gtRNAdb/Ensembl	tRNA	tRNA	158	99
Ensembl	tRNA_pseudogene	tRNA		
Ensembl	Mt_tRNA	tRNA		
Ensembl	Mt_tRNA_pseudogene	tRNA		
Ensembl	rRNA	rRNA	43	27
Ensembl	rRNA_pseudogene	rRNA		
Ensembl	Mt_rRNA	rRNA		
Ensembl	snRNA	snRNA	13	12
Ensembl	snRNA_pseudogene	snRNA		
Ensembl	snoRNA	snoRNA	10	6
Ensembl	snoRNA_pseudogene	snoRNA		
Ensembl	misc_RNA	misc_RNA	94	85
Ensembl	misc_RNA_pseudogene	misc_RNA		
Ensembl	lincRNA	misc_RNA		
Ensembl	scRNA	misc_RNA		
Ensembl	scRNA_pseudogene	misc_RNA		
Ensembl	pseudogene	misc_RNA		
RefSeq	CDS	misc_RNA		
RefSeq	INTRON	misc_RNA		
RefSeq	UTR	misc_RNA		
RefSeq	3FLANK	misc_RNA		
RefSeq	5FLANK	misc_RNA		
	unknown	unknown	80	56

as simple grid search, genetic algorithms or specialized methods such as VALP [Zien et al., 2000], we can optimize \mathcal{S} for many purposes, e.g. a median based hierarchical clustering that is supposed to separate all classes equally well would require a scoring system, that maximizes $\sum_A \text{median}(P^{\text{inner}}) - \text{median}(P^{\text{outer}})$.

Here we use only $P^{\text{outer}}(A)$ to test the null hypothesis, that an interval I without annotation is not from class A . We calculate an empirical p-value for each $\hat{s}(I, J)$, $J \in C^A$ from the right tail of $P^{\text{outer}}(A)$ and combine each of these $|C^A|$ p-values using Fisher's method [Fisher, 1970]. We then select the class with the smallest p-value.

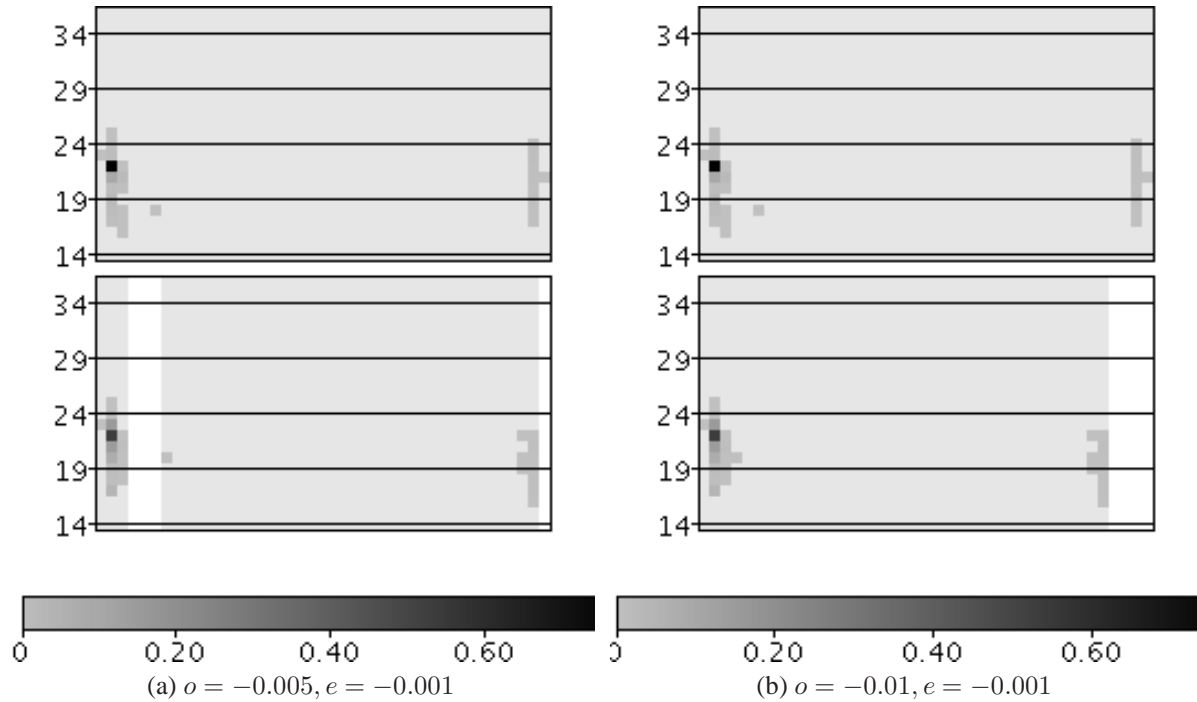


Figure 3.2: Freeshift alignment of hsa-mir-99b and hsa-mir-185. Pattern matrix frequencies are visualized in different shades of gray and white areas correspond to unaligned parts of the matrices. Note that with the gap cost function in 3.2a, the microRNA star start positions on the right parts of the matrices are correctly aligned, whereas slightly altered parameters in 3.2b erroneously move the necessary gap to the end, where it is not penalized.

3.5 Results

We applied our method to previously published Solexa sequencing data of human embryonic stem cells [Morin et al., 2008], where small RNAs of human embryonic stem cells (hESC) and embryoid body cells (EB) have been sequenced. We used Bowtie to align the trimmed reads to the human genome (hg19) obtained from the UCSC genome browser. We allowed no mismatches but did not restrict the number of loci a read can be aligned to. We identified intervals as described in the methods section ($t = 50, m = 1000$) and assigned them to the classes in table 3.1. We determined the normalized pattern matrices (see project website for graphical visualizations) and computed all pairwise ALPS similarities for various scoring systems.

First, we checked which choices of gap parameters make differences in the alignments of intervals. We considered the intervals of hsa-mir-99b and hsa-mir-185, that are both 5' donors (i.e. the mature microRNA originates from the 5' arm of the precursor), are expressed at similar levels (1892 and 2148 reads in EB, respectively) and have different loop lengths. Thus, a correct alignment must introduce a gap between the positions of the mature microRNA and the microRNA star in the sequence of column vectors of mir-185 (which has the shorter loop). If

38.3. Classification of ncRNAs using position and size information in deep sequencing data

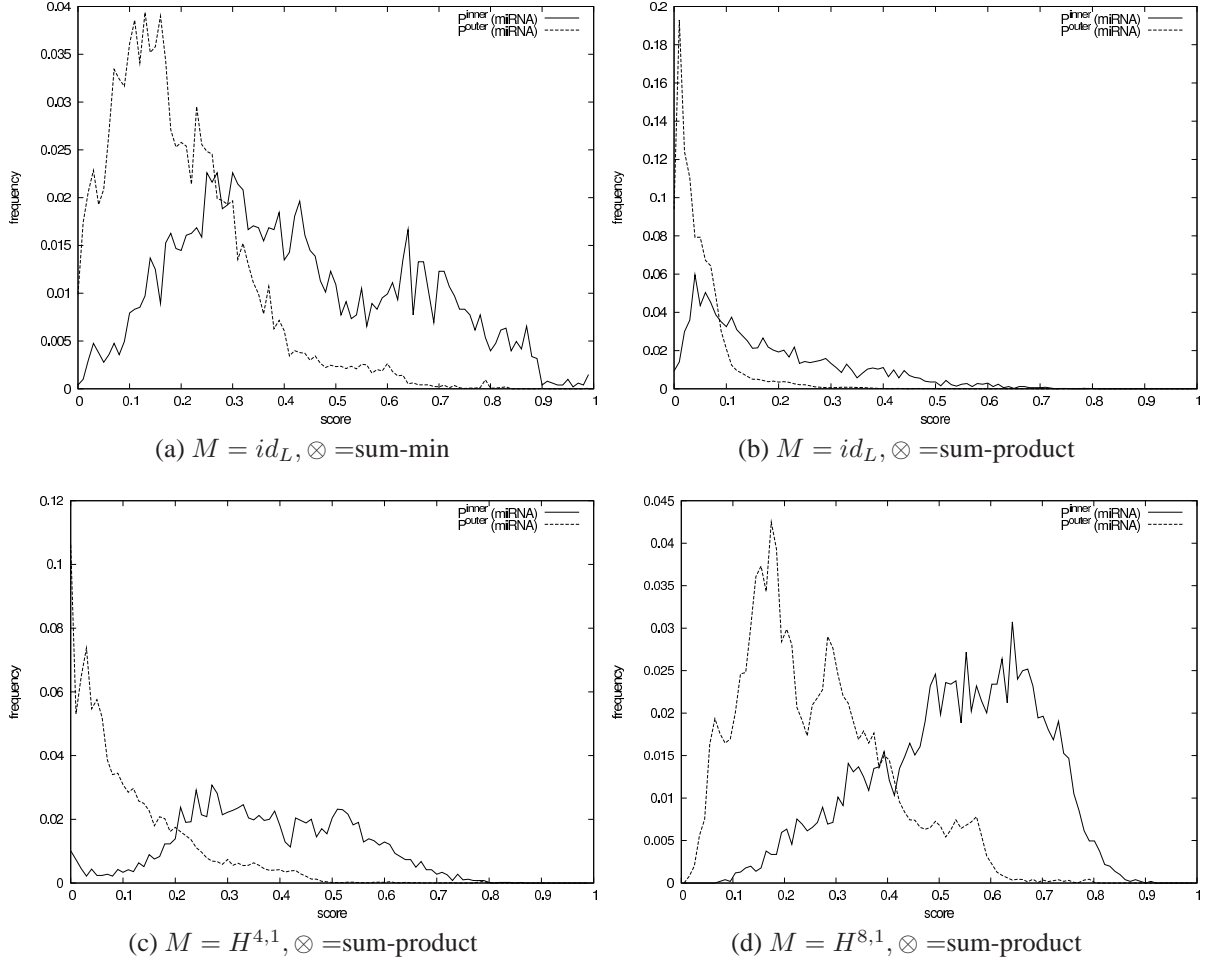


Figure 3.3: Inner and outer score distributions for microRNAs. The EB dataset is used and shown are scores for freeshift alignments with $o = -0.05, e = -0.01$. In all cases, the inner and outer distributions are significantly different, although not completely separated.

we calculate the optimal freeshift alignment using the hill matrix $H^{2,1}$ and min-product matrix multiplication, gap parameters of $o = -0.005$ and $e = -0.001$ are indeed able to produce a correct alignment (see Figure 3.2). We emphasize, that meaningful ranges of gap parameters are highly dependent on the other parameters and that automated parameter optimization could resolve these ranges.

A second theoretical consideration can be made by examining the inner and outer score distributions (see Figure 3.3). When the sum-min and the sum-product operator is used with the same matrix (the identity matrix), scores of the former naturally tend to be higher than scores of the later. If the identity matrix is replaced by $H^{4,1}$ or $H^{8,1}$, scores also tend to increase. For all parameter choices, it is apparent that inner and outer scores are significantly different, but their distributions are not completely separated. The outer distribution describes all ALPS similarities between pairs of intervals, one annotated as microRNA, the other not annotated as microRNA.

However, mirBase is not complete, and as a consequence, it is possible that the outer score distribution contains microRNA-microRNA scores, which can explain the elongated right tail of all P^{outer} . The inner distribution consists of all pairwise ALPS similarities of two intervals both annotated as microRNA. Especially when using $M = id_L$, many scores tend to be small, since only exact agreements in length are rewarded, and two mature microRNAs may have differing sizes. In addition, microRNA may be 5' donors or 3' donors or both mature and microRNA star are expressed at very similar levels. As a consequence, P^{inner} does not only contain overall high scores, but also scores indicating differing subclasses.

However, all these parameter choices are able to separate microRNAs from other ncRNAs, when we use all scores for classification. Using any aggregate statistics fails in many cases: If the maximal scores is used, a true microRNA may be too similar to an interval with unknown annotation, which is in fact a still unknown microRNA, leading to a misclassification. If one uses the minimum, the inner scoring is hampered by subclasses. Therefore, using all scores and a statistically robust method to combine them (such as Fisher's method) is necessary for reliable classification.

In order to assess whether ALPS is able to classify ncRNA reliably, we applied the following procedure: Each annotated interval I was removed from its cluster C^A and the described method was used to determine the class of I . Since we did not restrict the number of loci a read could align to, and many of the abundant ncRNAs are present in multiple copies in the human genome, we considered only scores $\hat{s}(I, J)$ where the genomic sequences of I and J did not contain common subsequences of length > 10 , i.e. no deep sequencing read has been counted in both intervals I and J . For all other scores, p-values were calculated and combined as described. We then calculated recall and precision for each class A separately as the number of intervals correctly assigned divided by the number of intervals originally belonging to C^A (recall) and divided by the number of intervals assigned to C^A (precision), respectively.

As indicated above, we tried various parameter combinations to classify ncRNAs. Since there are only very few unique snRNAs, snoRNAs and rRNAs, we only considered microRNAs and tRNAs for evaluation. Except for some obviously too extreme parameter combinations (e.g. too negative gap parameters for global alignments), the classification performance was remarkably stable with recall values of up to 98% at a precision of 60% for microRNAs (see Figure 3.4). These relatively low precision values in the microRNA class rise the question, whether our scoring tends to classify too many intervals as microRNAs. However, the classes *unknown* and *misc_RNA* are not excluded from our analyses, and nearly all of the intervals additionally assigned to the class microRNA originate from *unknown* and *misc_RNA* whose pattern matrix indeed is very similar to that of microRNAs. We predicted the secondary structures of the corresponding sequences using RNAfold [Hofacker et al., 1994] and some of them are indeed predicted to be able to fold into hairpins. Whether these reads really correspond to mature microRNAs, are degradation products or otherwise processed RNAs must still be elucidated, however.

Here, we applied our method only to abundant ncRNAs. This is inherent to the method as we have to estimate the distribution of read lengths per position for an ncRNA gene, which is only possible, if enough reads have been sequenced. Due to further development of current sequencing techniques, it will be possible to achieve more and more sequencing depth at lower costs and therefore, also low abundant ncRNAs will be represented by enough sequencing reads.

40 3. Classification of ncRNAs using position and size information in deep sequencing data

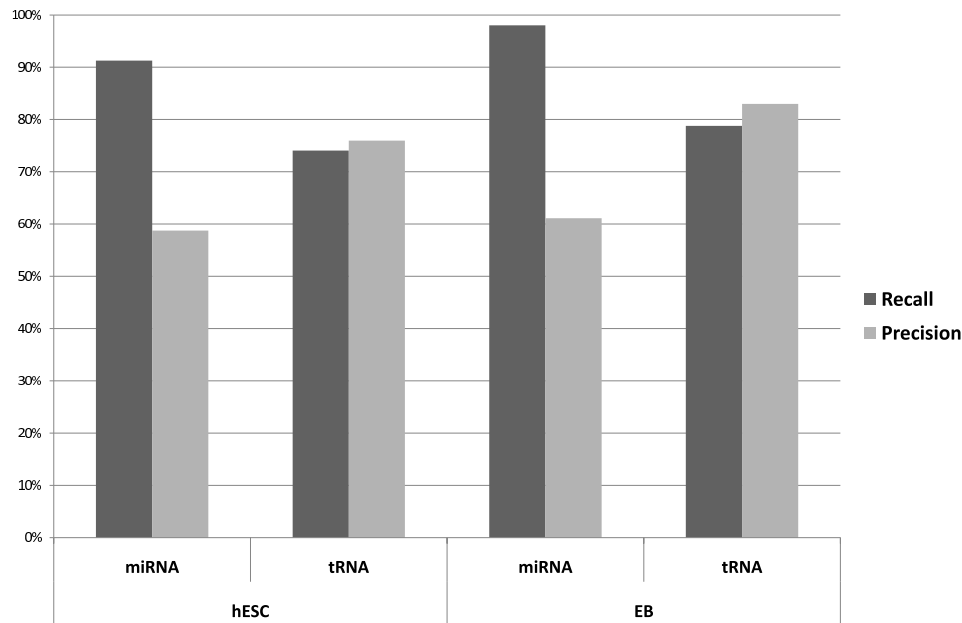


Figure 3.4: Evaluation for the scoring system ($H^{2,1}$, sum-min, -0.01 , -0.005 , freeshift). The recall for a class is the number of intervals correctly assigned divided by the number of intervals originally belonging to this class and the precision is the number of intervals correctly assigned divided by the number of intervals assigned to the class.

3.6 Discussion

Deep sequencing reads of ncRNAs follow very specific patterns regarding their length and positions with respect to their genes. Classes of ncRNAs are defined by their function and biogenesis and often share a common structure. Each of these can contribute to a biased distribution of reads on the ncRNA gene:

- Regulatory RNAs such as microRNAs, piRNAs or siRNAs are believed to recognize their targets by a short complementary region (seed). Therefore, for a proper function, the cell has to take care that the seed of these RNA classes is not shifted and, as a consequence, a wealth of deep sequencing reads starting at specific positions should be detectable. This can be observed in the pattern matrices computed for high-throughput sequencing data. The consideration of reads starting at adjacent positions allows to distinguish these ncRNA classes from degradation products of other abundant species.
- The specific pattern observed for longer ncRNAs like tRNAs (see Figure 3.1a) can possibly be explained by their degradation: Cleavage by RNases can be biased towards certain parts of the tRNA, which leads in the case of the cloverleaf structured tRNAs to a pattern of tRNA halves or quarters [Thompson and Parker, 2009]. Although some of these degradation products can be mistaken for e.g. a microRNA due to similar length,

the consideration of longer intervals and the distances between such subintervals can be used to separate these classes of ncRNAs. It has been observed, that degradation products of tRNAs are associated with RISC [Haussecker et al., 2010], which could explain microRNA-like read patterns of tRNAs. In spite of that, ALPS is able to separate these tRNAs from microRNAs.

- Patterns generated by microRNA biogenesis are obvious when looking at graphical representations of pattern matrices. In addition to the mature microRNA and the microRNA star, additional reads are present for some intervals. These can either be explained by degradation products or by additional drosha products that have been observed previously [Shi et al., 2009].

It is currently unclear which classes of ncRNAs exhibit characteristic patterns of deep sequencing reads, but the points discussed above indicate that in theory all ncRNA classes defined by a common function or biogenesis should have such a pattern and should therefore be amenable to classification by ALPS.

As indicated in Figure 3.2, exact distances between the start positions e.g. of the mature microRNA and the microRNA star within such patterns are not fixed and in plants, the microRNA hairpin is longer and even more variable than in animals. Allowing gaps in the alignment therefore enables ALPS to compute reasonable similarity scores for ncRNA classes, where such distances are highly variable. It is furthermore important to allow affine gap cost functions since linear gap costs tend to disrupt correct alignments. Gap parameters can be adjusted, such that single alignments become correct (e.g. as in Figure 3.2). However, we observed that classification accuracy in our test datasets is not heavily influenced by gap parameters. This is a consequence of the strong signal of the mature microRNA read that contributes in many cases enough score to the ALPS similarity to separate microRNAs from non-microRNAs. Thus, for classification of ncRNAs exhibiting such dissimilar patterns as in our test set, results are very robust and independent of the scoring system, i.e., a single reasonable scoring system can be used for classification of all ncRNAs in such a case.

If patterns for ncRNA classes are not as distinct as for the microRNAs and tRNAs in our test set, gap parameters and the matrix M can be tuned for proper classification. We described an approach to evaluate parameter sets based on the inner and outer score distributions and we note that already available methods to optimize other alignment based scoring systems e.g. for homology modelling and of protein (structure) alignment can directly be applied to ALPS.

We emphasize, that ALPS similarities should be calculated per experiment and comparisons across different datasets, generated in different labs with different protocols or even different sequencing platforms, should be performed with care. In the two datasets we used for validation, pattern matrices were highly concordant for intervals observed in both datasets. However, it is not clear how much technical bias is introduced into pattern matrices, i.e. pattern components, that are not due to biology but introduced by technical factors. Comparing pattern matrices of different protocols or sequencing techniques is subject for further studies.

3.7 Conclusion and Outlook

We developed an alignment based method, that allows to quantify the similarity of ncRNAs solely based on primary deep sequencing data by considering the position and length dependent patterns of reads aligned to short intervals on the genome. ALPS similarity rewards matching positions of reads of similar length in the two intervals. It can be computed efficiently and can be used to classify intervals of unknown function in various ways, one of which we have presented here.

ALPS only considers data, that is available by a deep sequencing experiment and makes no further assumption about the common secondary structure of an ncRNA class. Such a scoring system is important not only because the RNA secondary structure prediction is not always accurate, but also because some ncRNA classes may not even have a common secondary structure. As long as members of a class share a similar pattern of read lengths and positions, our method is able to detect it. For instance, there is no method available to accurately detect usRNAs [Li et al., 2009] in deep sequencing data in an automated manner. Since usRNAs are associated with RISC and their importance in post-transcriptional regulation has been shown, it is of great importance to provide a tool for their detection. Since they are characterized by their short length and fixed positions, ALPS similarity can be expected to identify them accurately in a deep sequencing experiment.

Our method can be used to support other, e.g. structure based, methods for the discovery of (specific) ncRNA classes by incorporating our similarity scores into the respective probabilistic model or machine learning scheme. As discussed, parameters of our scoring system can be finetuned in favor of any class of ncRNAs. In addition, if read lengths and positioning is also characteristic for subclasses, our scoring can be used to recover this hierarchy and for instance divide the class of microRNAs into the subclasses of 5' and 3' donors.

It has been suggested, that microRNAs are modified after maturation [Morin et al., 2008]. These modifications are detectable in a deep sequencing experiment, and if they are specific for microRNAs, incorporating them into a scoring system should further boost the identification of microRNAs. This can easily be incorporated into the calculation of the similarity score by extending the column vectors and defining appropriate matrices M . Even structural information could be integrated the same way.

We have shown that only considering positions and lengths of deep sequencing reads already allows to accurately identify abundant microRNAs and tRNAs in a large-scale dataset. Our scoring system was not biased towards the identification of a specific class of ncRNAs and as a consequence, we expect it not only to be useful for the classification of known ncRNA types, but also for novel classes, as long as they exhibit a characteristic pattern of deep sequencing reads.

Chapter 4

PARma: identification of microRNA target sites in AGO-PAR-CLIP data

Motivation: *The main idea of ALPS scores described in the previous chapter is to specifically exploit features of aligned sequencing reads and to find similar loci on the genome with respect to these features. Here, this idea is reused and developed further: Instead of computing similarities between pairs of loci, a general model of sequencing read features is built and all loci are classified with respect to the similarity to this model instead of computing all pairwise similarities. Furthermore, while the goal of ALPS scores was to classify ncRNAs in regular short RNA-seq data, PARma analyzes PAR-CLIP data in order to identify microRNA target sites as well as the microRNAs that target each site. Thus, due to the different nature of the data, the features used in PARma are different from the features in ALPS. Furthermore, based on close inspection of PAR-CLIP data using the data viewer described below, it was straight-forward to build a general model instead of all pairwise comparisons for valid microRNA target sites, while for short RNA-seq data, general models for the various ncRNA classes seemed not feasible. I applied PARma to new PAR-CLIP data generated by our collaboration partners as well as to published datasets. The results of these analyses are presented in chapter 6.*

Publication: *This chapter has been published in Genome Biology [Erhard et al., 2013a]. Here, I adapted the layout and made minor corrections to the text.*

My contribution: *I came up with the idea and the method, implemented the method, carried out evaluations and wrote the paper.*

Contribution of co-authors: *Lukas Jaskiewicz performed PAR-CLIP experiments. Lars Dölken contributed ideas and helped to revise the manuscript. Ralf Zimmer supervised the work and helped to revise the manuscript*

4.1 Abstract

AGO-PAR-CLIP is a high-throughput method to identify target sites of microRNAs based on immunoprecipitation (IP) of the RNA induced silencing complex (RISC) and deep sequencing Argonaute (Ago)-protected RNA fragments crosslinked to Ago. This approach provides clusters of reads spanning 30-50 nucleotides containing the microRNA binding sites. The identity of the microRNA binding in each cluster is a priori not clear and must be revealed by matching the correct microRNA seed sequence to the cluster sequence, which is not a trivial task.

Specific characteristics of PAR-CLIP data can be utilized to aid this problem, most notably, frequent T to C conversions that are indicative for crosslinking sites. We utilize these and additional features to accurately determine the seed site. Our method, PARma, consists of two main components: A generative model incorporates PAR-CLIP specific features to compute likely seed site positions and the novel pattern discovery tool *kmerExplain* estimates seed activity probabilities based on the likelihood inferred by the model.

The final PAR-CLIP model is in agreement with known binding mechanisms of microRNAs and with structural knowledge of AGO and many active k-mers correspond to seeds of expressed microRNAs. Based on the analysis of differential PAR-CLIP data from both a publicly available dataset as well as from a new dataset, we show that PARma is more accurate than existing approaches in terms of correct seed assignments.

PARma is freely available from the project website <http://www.bio.ifi.lmu.de/PARma>.

4.2 Introduction

MicroRNAs have emerged as important post-transcriptional regulators in all known multicellular organisms. These 20-24 nucleotide (nt) long RNA molecules play roles in development, tumorigenesis and viral infection [Bartel, 2004]. Generally, they bind to 3' UTRs of their target transcripts inhibiting translation or inducing degradation of the target mRNA [Bartel, 2009]. Neither the exact mode of binding nor the mechanisms of downregulation are completely understood and are under heavy debate [Djuranovic et al., 2011; Eulalio et al., 2008; Guo et al., 2010; Kozak, 2008; Mishima et al., 2012]. It is believed that microRNAs recognize their target sites using only a small portion of bases at their 5' end called the seed [Wee et al., 2012] and that other factors such as additional base pairing at the 3' end [Bartel, 2009], target site accessibility [Kertesz et al., 2007], target site location, AU content around the target site contribute to recognition [Grimson et al., 2007]. These factors, as well as evolutionary conservation of target sites (in case of conserved microRNAs) have been used to predict target sites of microRNAs [Friedman et al., 2008; Krek et al., 2005]. However, all known prediction methods are hampered by a huge number of false positives and false negatives [Ritchie et al., 2009]. Recently, several high-throughput assays have been developed which allow accurate identification of microRNA targets (reviewed in Thomson et al. [2011]).

Immunoprecipitation (IP) of the Argonaute (AGO) protein, the major component of the RNA induced silencing complex (RISC), allows the identification of microRNA mediated recruitment of hundreds of different transcripts to the RISC. Target mRNAs of microRNAs co-precipitate

with AGO and can thus be identified either using microarrays (RIP-Chip) or Next-Generation-Sequencing (RIP-seq) [Easow et al., 2007; Beitzinger et al., 2007; Hendrickson et al., 2008; Karginov et al., 2007; Landthaler et al., 2008; Dölken et al., 2010]. However, these RIP experiments only give information about target genes or transcripts and neither about the precise location of target sites nor the actual microRNA targeting these sites. As a remedy to that, novel techniques including HITS-CLIP, iCLIP and PAR-CLIP have been developed. Before the IP, RNA is cross-linked to proteins using UV light, which allows then to determine the precise location of the target site by deep sequencing of cross-linked RNA after digestion of non-cross-linked RNA [Chi et al., 2009; König et al., 2010; Hafner et al., 2010]. Still, the actual microRNA binding at these sites have to be determined.

Both techniques, RIP and CLIP, need specialized bioinformatic analysis methods. RIP is very similar to standard gene expression experiments and, thus, advanced analysis methods are readily available. In addition to these standard approaches, in a recent paper, we described additional algorithms which need to be employed to consider and cope with the characteristic features of RIP data [Erhard et al., 2013b]. In contrast, CLIP data are more complex: First, short sequencing reads must be aligned to the genome or transcriptome and then clustered [Chi et al., 2009; König et al., 2010; Hafner et al., 2010]. True target sites have to be identified among all clusters and the specific microRNA targeting each site has to be determined. Depending on the exact experimental protocol, true target sites may look quite distinctive: While for HITS-CLIP, narrow peaks in the read coverage are expected [Chi et al., 2009], iCLIP clusters show specific read start positions [König et al., 2010] and PAR-CLIP clusters are characterized by T to C conversions [Hafner et al., 2010]. Here, we focus on PAR-CLIP, a technique that has been used by several groups to identify microRNA target sites [Hafner et al., 2010; Gottwein et al., 2011; Lipchina et al., 2011; Skalsky et al., 2012].

In their original PAR-Clip paper, Hafner et al. [Hafner et al., 2010] used several manually chosen parameters to define target sites (e.g. at least two distinct conversion positions per cluster and at least 5 sequencing reads). They recognized that the region downstream of the main conversion site is enriched for sequences complementary to the seeds of top expressed microRNAs.

PARalyzer is a software package specifically designed to define RNA binding sites from PAR-CLIP data. Reads are first clustered and filtered using similar parameters as Hafner et al. Then, conversion and non-conversion distributions are computed by counting the respective events and employing kernel density estimation along each cluster. All positions with a higher conversion than non-conversion density are considered target sites and surrounding sequences are submitted to a standard motif discovery tool that uses linear regression to determine microRNA seed sites enriched among clusters with many conversion events [Corcoran et al., 2011].

There are several open points in PAR-CLIP data analysis: First, it is unclear which microRNAs should be taken as starting point for searching seed sites in PAR-CLIP clusters. In all published studies, the top N microRNAs according to microRNA read counts in the PAR-CLIP experiment or an additional experiment are taken. However, read counts provide a potentially strongly biased estimate of microRNA expression levels [Raabe et al., 2011; Linsen et al., 2009]. In addition, it is unclear how many miRNAs should be used. Finally, it may not be sufficient to only consider known microRNAs: First, there are indications that there are many still unknown microRNAs [Ladewig et al., 2012] and second, not only microRNAs (as defined by their maturation pathway)

may be associated with AGO and used for target recognition, but there may be other pathways that lead to the incorporation of small RNAs into RISC [Haussecker et al., 2010; Cheloufi et al., 2010; Yang et al., 2010; Cifuentes et al., 2010; Taft et al., 2009].

Second, the specific information given by the PAR-CLIP experiment is only partially exploited: In the PAR-CLIP protocol, RNase T1 is used to digest RNA, which cleaves specifically after guanine [Pace et al., 1991]. This information could be used to exclude seed sites spanning read start or end positions under the assumption that these sites are protected from digestion by the microRNA. Also, it is known that positions in the mRNA bound to the microRNA cannot be efficiently cross-linked and thus, seed sites spanning a cross-linking site could also be excluded [Hafner et al., 2010]. Currently, there is no method available that directly uses the information from RNase cleavage sites or conversion sites for the discovery of motifs or the assignment of seed sites.

Third, there is no scoring system available that has been demonstrated to reliably identify clusters or assigned microRNAs.

Here, we present a method to address these aspects: PARma seeks explanations for the presence of each identified PAR-CLIP cluster. Here, an explanation is a k nt long sequence (k -mer) within a cluster that corresponds to the seed of the microRNA binding this site. PARma explains each PAR-CLIP cluster by a k -mer that is (a) explaining multiple clusters with high probability and (b) matching a generative model for the experimental data (i.e. the data observed in the experiment is likely to be generated by a microRNA binding at the determined position). The determined k -mer can identify respective microRNA families that are characterized by a seed matching the k -mer. The model is able to score each k -mer in a cluster according to the observed conversions and RNase cleavage sites. Parameters as well as k -mer activity probabilities are estimated in an iterative manner. The model assigns the most probable seed to a PAR-CLIP cluster, to score clusters according to the confidence of being a true microRNA target site and also to score the confidence of the assignment of the correct seed.

Differential PAR-CLIP data are used to evaluate our methods: When paired PAR-CLIP datasets with microRNAs that are known to be present only in dataset A and not in B are analyzed, target sites (PAR-CLIP clusters) of these microRNAs should only be present in dataset A. We used our own PAR-CLIP datasets of the two B-cell lines DG75 and BCBL1, of which only the latter is infected with Kaposi's sarcoma-associated herpesvirus (KSHV), a herpesvirus encoding 25 mature microRNAs. In this data, we expect the viral microRNAs and hence its targets only to be present in the infected cell line. We also repeated our evaluations using a published dataset of Epstein-Barr-Virus (EBV; encoding 44 mature microRNAs) positive and negative cell lines [Gottwein et al., 2011].

4.3 Results

4.3.1 PARma overview

We developed a complete workflow for the analysis of PAR-CLIP data (see Figure 4.1). The main steps are (a) mapping of the sequencing reads to reference sequences, (b) detection of read

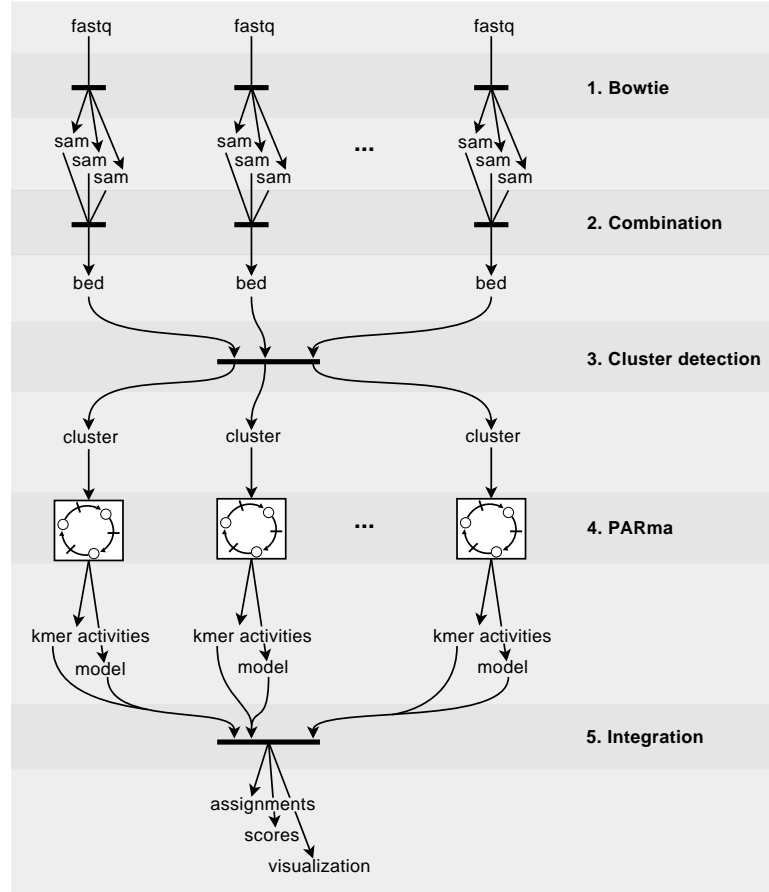


Figure 4.1: PARma overview. The PARma workflow starts with the raw data from PAR-CLIP experiments (replicates or different conditions), i.e. several fastq files containing sequencing reads. First, we utilize Bowtie [Langmead et al., 2009] to align these reads to multiple reference sequences such as the human genome and transcriptome or viral genomes resulting in several sam files, one for each fastq file and reference sequence. Second, for each read from each experiment we identify all optimal alignments in terms of mismatches, considering T to C conversions as matches, and map transcriptomic reads that span splice junctions to the genome. Third, possible target sites of microRNAs are identified by clustering reads from all datasets simultaneously. The clusters including additional annotations such as the number of conversions and cleavages per position are written to separate files for each experiment individually. The cluster detection implements a splitting procedure to identify target sites with overlapping reads and is able to handle target sites that span splice junctions. Fourth, for each dataset, the core PARma component estimates a generative model for the data and k-mer activity probabilities using kmerExplain in an iterative manner (see also Figure 4.3). Fifth, the models and the activity probabilities are used to score clusters and to assign the most probable microRNA. Target sites with various annotations such as gene ids are written to tabular files that can be further analyzed and visualized.

clusters corresponding to target sites, (c) estimating a model that represents characteristic features of PAR-CLIP data and microRNA (seed) activities and (d) the final assignment of microRNAs to target sites and their scoring using the derived model. Furthermore, we developed a tailored, web-based visualization for PAR-CLIP data, that helped us during the development of PARma and can be used to manually investigate specific target sites (see Figure 4.2).

The central idea of PARma is that microRNAs binding a target site will generate specific data in a PAR-CLIP experiment (conversion positions and RNase T1 cleavage sites, see Figure 4.2a). Thus, given experimental data and a model representing these features, it is possible to infer the binding site that has generated these data with the highest likelihood. Additionally, given the experimental data and the correct binding sites, it is straight-forward to infer the model parameters. Thus, we are facing a chicken-or-egg dilemma: If we knew the binding sites we could infer the model, and if we knew the model, we could infer the binding sites. In PARma, this is resolved using an iterative procedure (see Figure 4.3). We start by computing statistically overrepresented k-mers in clusters and take these as initial estimates for the correct binding sites. Then, we infer model parameters and iteratively refine all estimates until convergence.

During these iterations, seed activity probabilities are estimated, corresponding to the likelihood-weighted number of target sites. Importantly, it is possible - but not necessary - to specify an a-priori set of allowed microRNAs. This is a highly desirable feature since in general it is unknown, which microRNAs are active in an experiment and the read count of the microRNAs themselves in the PAR-CLIP experiment or an external sequencing experiment is only a weak proxy for their activity, as shown below.

In the final output of PARma, for each cluster the most probable seed is assigned together with a cluster score (Cscore) and a microRNA assignment score (MAcore). The Cscore indicates how well the observed data (conversions and RNase cleavage sites) fit the model without considering the k-mer probability and therefore indicates, whether an observed clusters is indeed a true microRNA target site. The MAcore corresponds to the confidence of the assignment, i.e. whether there are other active k-mers in the cluster that also match the observed data well.

4.3.2 Cluster detection

After read mapping (see Methods), the first main step of PAR-CLIP data analysis is to identify clusters of reads corresponding to target sites. Overall, we use a similar procedure as has been used previously with a few but important modifications:

First, PARma is able to search for clusters using multiple datasets simultaneously. This not only increases sensitivity, but also provides a straight-forward way for a differential analysis of target sites, since it is not necessary to identify corresponding clusters from different experiments afterwards. During the cluster identification, clusters are determined for all datasets simultaneously, and each cluster is quantified for each dataset.

Second, the original definition of PAR-CLIP clusters (i.e. target sites) by Hafner et al. [2010] involved a single linkage clustering of overlapping reads. However, we observed several cases where such a procedure tends to link multiple target sites into a single cluster due to few spurious reads that connect two obviously distinct clusters (see Figure 4.4a for an example). Such cases are relatively frequent (see Figure 4.4b) and may be of special interest: For instance, there are

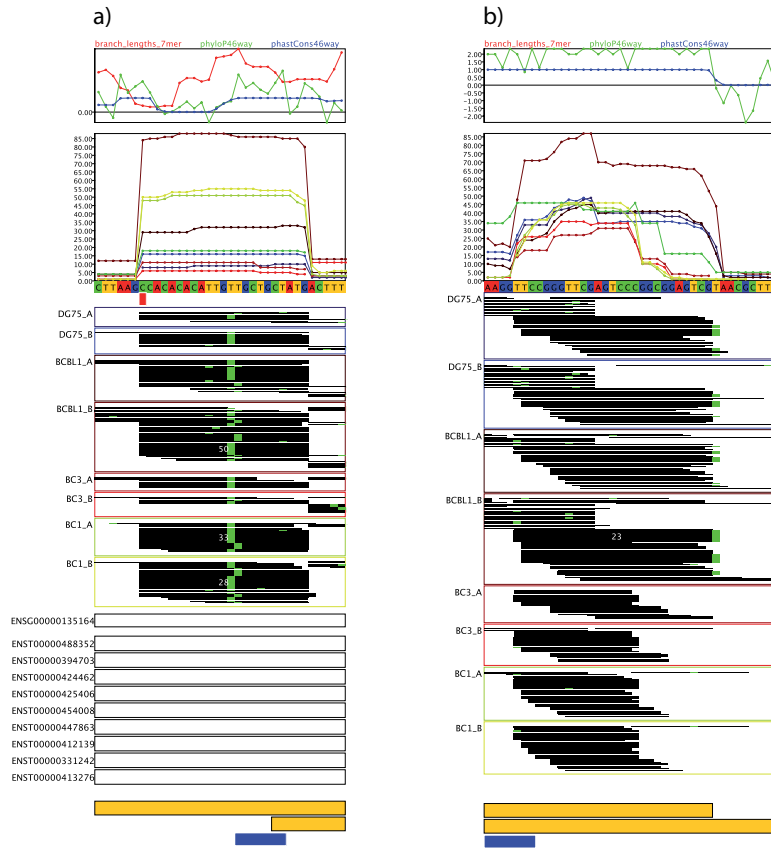


Figure 4.2: PAR-CLIP data viewer. From top to bottom both panels show conservation scores (branch lengths of 7-mers as described by Friedman et al. [2008] and the widely used phyloP [Pollard et al., 2010] and phastCons [Siepel et al., 2005] scores, all computed for the 46-way vertebrate multiple alignment obtained from the UCSC genome browser [Dreszer et al., 2012]), the read coverage in each experiment and the genomic sequence of cluster. Below the sequence, SNP positions according to the 1000 genomes project are indicated in red (here only in Figure 4.2a) and the actual sequencing reads are shown as black bars for each of the experiments. Mismatches are color-coded as in the genomic sequence on the top (i.e. in both clusters, there are T to C conversions only). The height of the bar directly corresponds to the read count in the PAR-CLIP experiment up to a count of 15 reads and more than 15 reads are indicated in white. Ensembl genes and transcripts are shown below the reads (here only in Figure 4.2a), together with PAR-CLIP clusters in yellow and seed site assignments in blue. In Figure 4.2a an experimentally validated targets site of hsa-miR-15 in the 3'UTR of *DMTF1* is shown. It illustrates the characteristic features of many valid target sites (see main text). Interestingly, there is also a known SNP (red box) in proximity to the seed site. Figure 4.2b depicts an intergenic (i.e. there are no Ensembl genes or transcripts) cluster that does not show these characteristics. Additionally, it does not contain a microRNA seed site nor any overrepresented 7-mer according to PARma. The validated cluster has Cscore and MAScore > 0.9, whereas for the intergenic cluster, both scores are 0.

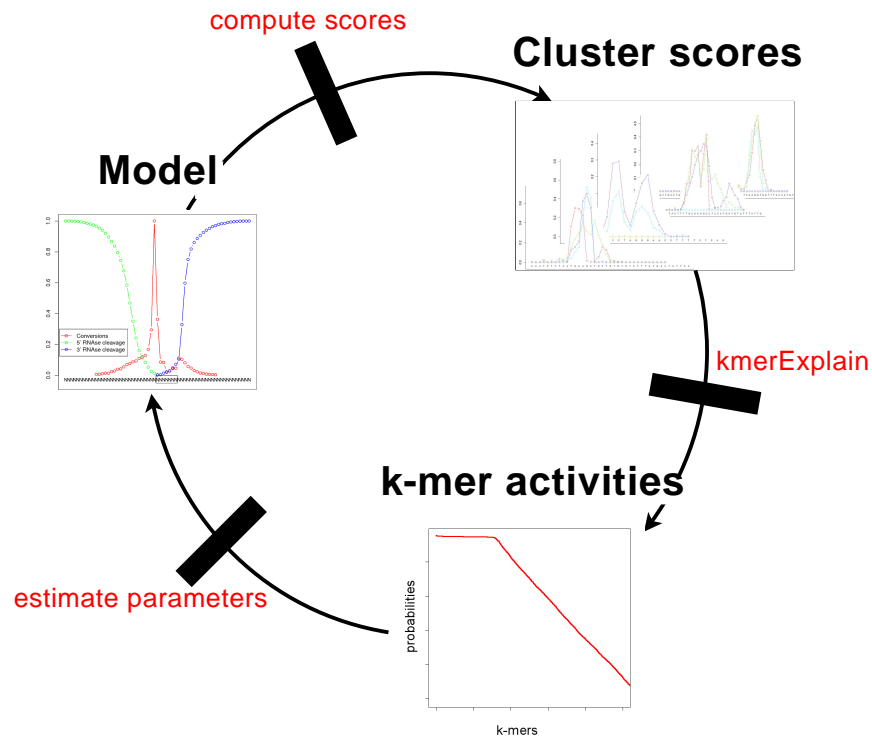


Figure 4.3: Illustration of the PARma procedure. PARma is an iterative algorithm that repeatedly executes three steps: Based on a current model of PAR-CLIP characteristics (left; see also Figure 4.6), scores are computed for each position in each cluster expressing the likelihood that the cluster is explained by the activity of the k-mer at this position (top right; see also Figure 4.7). These scores are fed into kmerExplain as prior probabilities, which then estimates k-mer activity probabilities using an EM algorithm (bottom). These k-mer activities in conjunction with data from the PAR-CLIP experiment (T to C conversions, RNase cleavage sites) are used to estimate the parameters of the PAR-CLIP model. We start this procedure by running kmerExplain on uniform scores and end it as soon as the model converges.

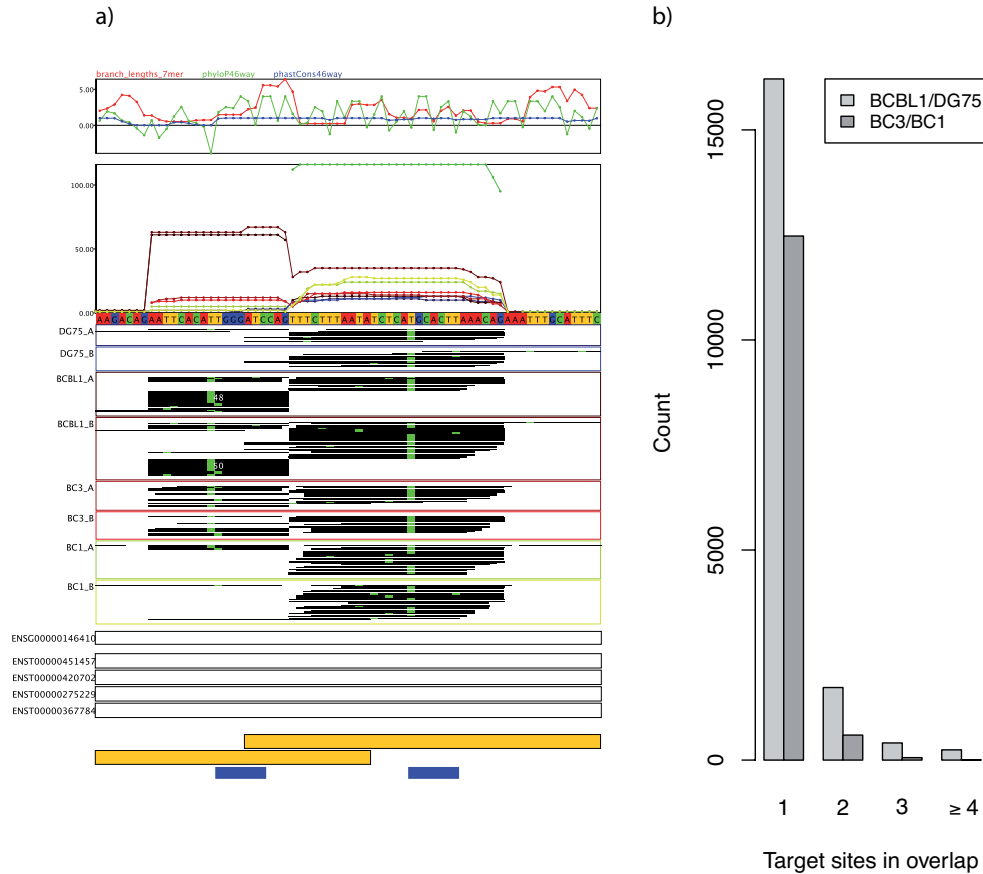


Figure 4.4: Overlapping PAR-CLIP clusters. In Figure 4.4a two target sites are shown that would fall into the same cluster by the definition of Hafner et al. [2010] only because in the two DG75 replicates as well as in the second BCBL1 replicate a few random reads from the right target site overlap the left target site. Our cluster definition splits all reads into two overlapping clusters (see the yellow boxes on the bottom. PARma rates both clusters with high Cscores (> 0.6 and > 0.9 for the left and right cluster, respectively) and assigns the KSHV microRNA kshv-miR-K12-7 to the left and the human microRNA hsa-miR-519 to the right cluster with MAScores > 0.9 in both cases. There is additional evidence that both assignments are correct, since the left cluster has reads only in KSHV positive cell lines (BCBL1, BC1 and BC3) whereas the right cluster contains reads in all experiments. Figure 4.4b illustrates that there are hundreds of such cases in both experiments.

cases known, where viral microRNAs bind to sites in close neighborhood to target sites of human microRNAs [Nachmani et al., 2010]. Missing individual clusters due to overlapping reads would be detrimental to such an analysis. Thus, we devised a cluster splitting procedure, that is able to effectively detect such cases.

And third, we align PAR-CLIP reads to the transcriptome as well as the genome. Transcriptomic reads are then mapped to genomic coordinates and may therefore produce spliced reads. These

are properly respected during cluster detection, i.e. PARma is able to detect target sites spanning exon-exon junctions. In previous studies about AGO-PAR-CLIP data [Hafner et al., 2010; Gottwein et al., 2011; Lipchina et al., 2011; Skalsky et al., 2012], this has not been considered, probably missing several highly interesting target sites. Indeed, in the datasets we analyzed, 22.4% of all clusters in the coding region of transcripts span splice junctions (about 6% of all clusters).

4.3.3 Generative model

The novel feature in PAR-CLIP (in comparison to other CLIP protocols) is the usage of the uridine analogue 4-thiouridine which is not read as U but as C during cDNA synthesis following its cross-linking to proteins [Hafner et al., 2010]. Thus, T to C mismatches of aligned sequencing reads are characteristic for cross-linked sites and, therefore, for contacts of the examined protein with RNA. Since RNase T1 is used in the PAR-CLIP protocol, which cleaves specifically downstream of guanine, it is of importance where sequencing reads start and end. It is important to note that in most cases, the RNase products are shorter than the number of sequencing cycles (36 for the data of Gottwein et al. [2011] and 50 for our data). Therefore, in these cases the complete RNA fragments are known.

Visual inspection of these features for known target sites of microRNAs using our PAR-CLIP data browser (see Figure 4.2) showed several characteristics of these targets sites that go beyond the characteristics of individual PAR-CLIP sequencing reads (see Figure 4.2a): In most cases, there is a main cross-linking site and $\geq 60\%$ of all conversions in the cluster belong to this site, a fact that has been recognized before [Hafner et al., 2010]. In addition, this main cross-linking site tends to lie in the center of most sequencing reads and T sites upstream tend to be cross-linked more often than T sites downstream of the main site. Another well-established feature is the position of seed sites preferentially downstream of the main cross-linking site. Finally, in addition to these main cross-linking sites, there are main RNase cleavage sites with specific locations, one ~ 10 to ~ 20 nt upstream of the seed site, the other usually immediately downstream of the seed site. While the upstream cleavage site often skips several G sites, the downstream site is in most cases immediately after the next G.

To formally represent these features, we developed three independent probabilistic models, the conversion model and the upstream and downstream cleavage models. Given the position of a seed site and the positions of uridines or guanines, respectively, each model is able to predict where and how many conversions or cleavages, respectively, would be generated by a PAR-CLIP experiment. By comparing the predicted to the measured data, we compute a likelihood for each possible seed position within a cluster. Specifically, the conversion model would generate many conversions directly upstream of the seed position (given there is a uridine), and almost no conversions within the seed. Thus, such a position would receive a high score only if this is indeed observed in the experiment.

Model parameters, e.g. how many conversions are expected for each uridine within a cluster, are directly learned from the data in a per-experiment manner using robust parameter estimation techniques. Doing this for each dataset individually is important, since experimental conditions

may be slightly different between experiments, potentially leading to slightly different data per cluster.

4.3.4 KmerExplain

KmerExplain optimizes a probabilistic model that requires that each target site is targetted by a single microRNA family, i.e. each cluster must be explained by a single k-mer (i.e. microRNA seed). There are two conditions for the explaining k-mer implicated by the model: First, its position in the cluster has to match the generative PAR-CLIP model, i.e. the given data (conversions and cleavages) are likely to be generated by a seed matching to this positions. And, second, the k-mer is likely to be active, i.e. there are many instances where this k-mer explains a cluster. The model is fitted with an EM algorithm.

4.3.5 Seed activities

We applied PARma to a previously published PAR-CLIP dataset consisting of two replicates for each of the two B-cell lines BC3 and BC1, as well as to our own PAR-CLIP data of two replicates for each of the two B-cell lines DG75 and BCBL1. First, we analyzed the correlation of microRNA expression as measured by its PAR-CLIP read count and its activity as measured by the number of assigned target sites.

Even if it is true that the top 100 expressed microRNAs may explain $> 50\%$ of the clusters by a 6-mer seed, the overall correlation between the microRNA expression and the number of corresponding target sites is poor (see Figure 4.5). This is a general observation, independent of how microRNAs have been assigned to the clusters (a variety of options have been explored: all or a random seed site in the complete cluster, the first or a random seed inside the cluster but downstream of the main cross-linking sites, using the top 40, 100 or 200 microRNAs, 6-mer or 7-mer seeds). The poor correlation may be a consequence of sequencing artefacts known to substantially bias expression estimates of microRNAs [Raabe et al., 2011; Linsen et al., 2009].

In addition, we and others proposed that not only microRNAs may enter the RISC pathway, but there may be other maturation pathways producing small RNA molecules that could act analogously to microRNAs in RISC [Haussecker et al., 2010; Cheloufi et al., 2010; Yang et al., 2010; Cifuentes et al., 2010; Taft et al., 2009; Ladewig et al., 2012; Erhard and Zimmer, 2010; Maute et al., 2013]. Furthermore, even if only the 7-mer seeds of the top 40 microRNAs are used and seed sites are only considered when downstream of the main cross-linking site, there are hundreds of clusters where two or more seeds match. Necessarily, this issue becomes more severe, if more than 40 microRNAs or all seed sites within the cluster are used (see Figures 4.5b and 4.5d).

Taken together, these facts suggest to abandon the paradigm of taking the top N expressed microRNAs as candidate regulators for PAR-CLIP clusters. Therefore, we designed PARma to identify k-mers among all possible 4^k k-mers that are explaining multiple clusters with high probability. Furthermore, they need not only to explain multiple clusters, but their positions must be in agreement with the model that is learned from the data of all clusters.

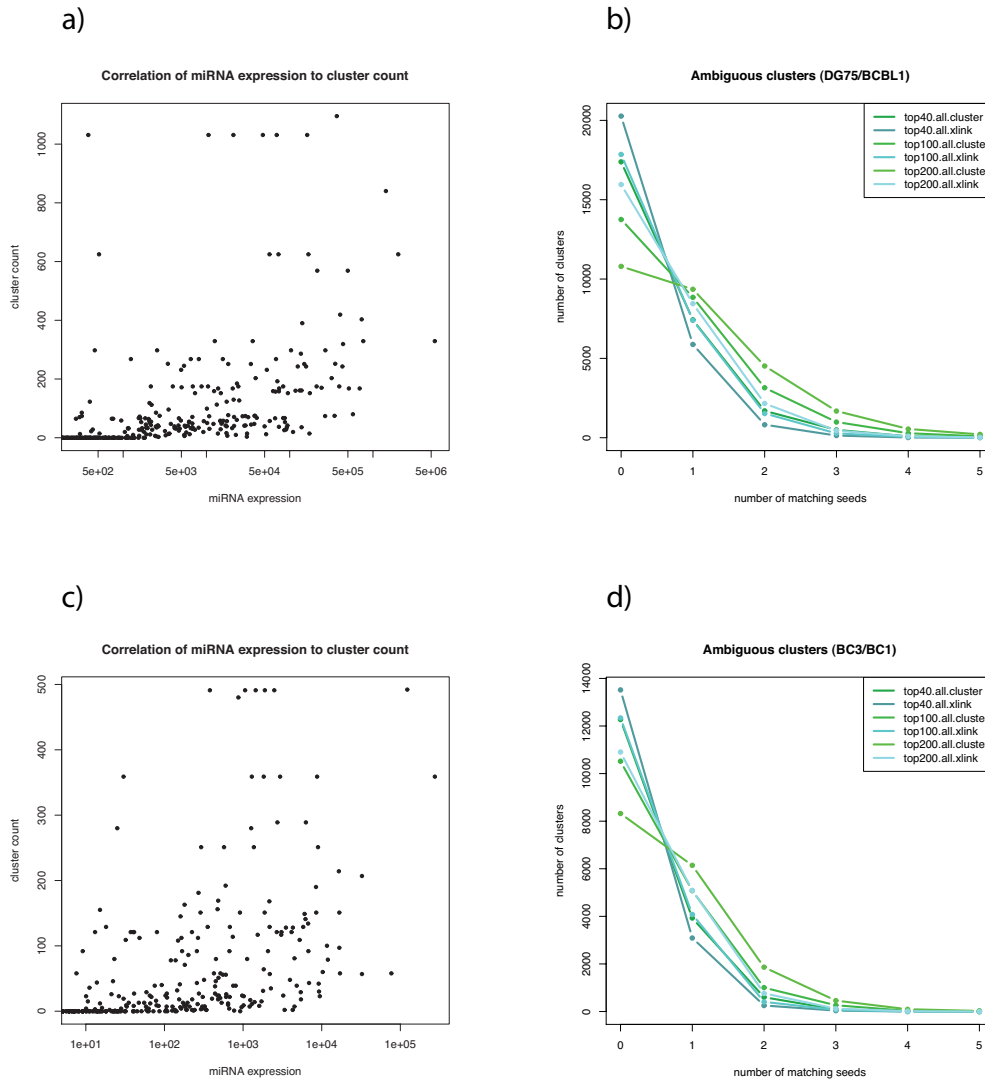


Figure 4.5: Correlation of microRNA expression to the number of assigned clusters. Here, microRNAs have been assigned to a cluster, when they are among the top 200 expressed microRNAs and match the first seed site downstream of the main cross-linking site. Neither in the BCBL1 PAR-CLIP data (Figure 4.5a) nor in the BC3 PAR-CLIP data (Figure 4.5c) any correlation is recognizable. Figure 4.5b and Figure 4.5d illustrate how many 7-mer seeds match to clusters, when the top 40, 100 und 200 microRNAs are considered and when seeds are searched in the whole cluster (*all*) and only downstream of the main cross-linking site (*xlink*). Even the strictest assignment (top 40 *xlink*) leads to a considerable amount of about 1000 ambiguous clusters in both datasets and at the same time to about 80% unassigned clusters. The fraction of unassigned clusters drops below 50% when the top 200 microRNA seeds are searched in the whole cluster but with the cost of having thousands of ambiguous assignments.

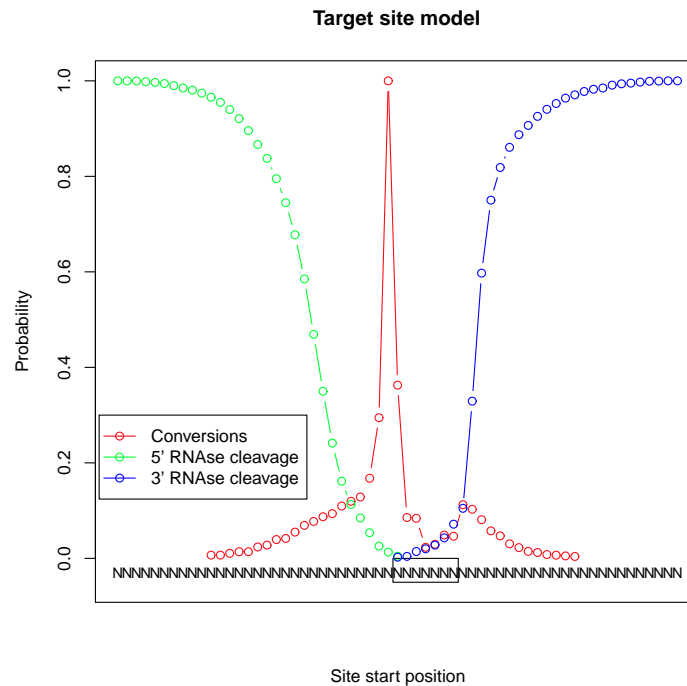


Figure 4.6: PARma model for replicate A of the DG75 experiment. The conversion model in red represents the conditional propensity that a base gets cross-linked given there is a uridine at the corresponding position. Note that the propensity is only known up to a constant factor and arbitrarily scaled to a mode of 1. The blue and green lines illustrate the 3' and 5' cleavage models, respectively. These correspond to the conditional probabilities that the RNase T1 cleavage site is at a certain position or closer to the seed site given that there is a guanine. The model shows that the observations made for a few visually inspected validated target sites are also true globally for many clusters.

4.3.6 Inferred models

Next, we analyzed the generative model that is estimated by PARma. In Figure 4.6, the model for replicate A of DG75 is illustrated. It indeed reflects the above mentioned observations: The conversion model indicates the expected ratios of conversions around the seed site for all positions where a T is located: For instance, if there is a T immediately upstream of the seed site and a T immediately downstream, the expected ratio of conversions is about 10:1. Furthermore, the first position in the seed site also seems to get cross-linked with relatively high frequency (for an example see also Figure 4.2a).

The models from Figure 4.6 are in agreement with knowledge about microRNA target recognition [Bartel, 2009]: A canonical microRNA binding site consists of a seed site complementary to the microRNA seed (bases 2 to 7 or 2 to 8), often base 1 is opposite of an A and often there is additional basepairing of the microRNA 3' end after a small loop. Thus, the seed site itself

may be protected from cross-linking by the seed, bases immediately upstream of the seed are accessible and further upstream bases may also be protected by the microRNA 3' end to some extent.

Furthermore, the model also agrees with known structural information of AGO2 [Schirle and MacRae, 2012]: MicroRNA bases 2 to 6 are solvent exposed and there is a distinct kink separating bases 6 and 7, which may be resolved by conformation changes of AGO [Schirle and MacRae, 2012]. These conformation changes may be a reason for the relatively high cross-linking probability of the first position of the seed site. Another explanation is that PARma may find several instances of 7-mer-m8 seed sites (pairing of bases 2 to 8) as well as 7-mer-A1 seed sites (pairing of bases 2 to 7 plus an A opposite of base 1). The first base of the identified k-mer may therefore be opposite of base 7 or 8 of the microRNA, and, therefore, may or may not be accessible for cross-linking.

As described above, all three submodels can be used to compute a score for each possible seed site position within a cluster. The conversion score (see Figure 4.7a for the cluster in Figure 4.2a) indicates that either immediately upstream or downstream of the main cross-linking site are likely positions for a seed site: The downstream position is obvious, the upstream position however is also probable, since further upstream there is no T that could get cross-linked. Figures 4.7b and 4.7c illustrate that the seed position is restricted to a small part of the cluster due to the clear 5' and 3' RNase cleavage sites. In addition, based on the estimate of kmerExplain, the k-mer *TGCTGCT* (see Figure 4.7d) is highly active and indeed corresponds to the 7-mer-m8 seed site of the in B-cells highly expressed miR-15/16 family. Hence, PARma is able to predict the corresponding position with high confidence, which is indeed an experimentally confirmed target site of miR-15a [Kiriakidou et al., 2004].

Although the PAR-CLIP protocol is rather stringent and thus provides reasonably pure AGO complexes, other RNA-protein interactions of co-purified proteins or abundant cellular proteins may be responsible for cross-linked and protein-protected RNA fragments, giving rise to non-AGO PAR-CLIP clusters. The model we developed also allows computing a cluster score (Cscore) indicating the likelihood by which a given cluster actually represents a microRNA binding site, i.e. how well the observed data (conversions and RNase cleavage sites) fit the model without considering the k-mer probability. The microRNA assignment score (MA score) indicates whether there are other overrepresented k-mers in the cluster that also match the observed data well. The experimentally confirmed target site in Figure 4.2a has Cscore and MA score of 0.9608 and 0.9777, respectively, whereas the cluster in Figure 4.2b has a Cscore of 0, indicating that there is no position where conversions and RNase cleavage sites agree.

4.3.7 Evaluation using differential PAR-CLIP

We evaluated PARma against PARalyzer and the standard approaches of assigning seeds of the top N microRNAs (for $N=40, 100$ and 200) when they are in the cluster (*cluster*) or downstream of the main cross-linking site (*xlink*) and either assigning every seed (*all*) or a *random*/the *first* seed (for *cluster* and *xlink*, respectively), when there are multiple seeds present. For the evaluation, we exploit a unique feature of the datasets we used: In our own data, only the cell line BCBL1 and not DG75 is infected by Kaposi's sarcoma-associated herpesvirus (KSHV), which

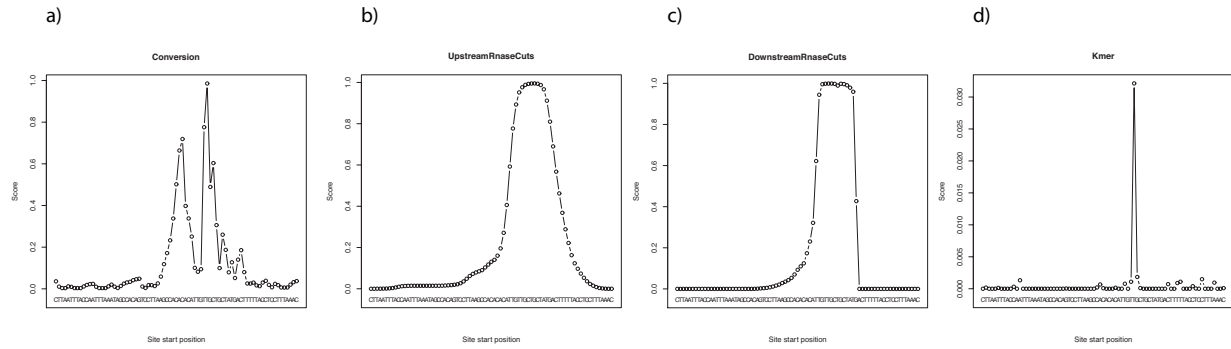


Figure 4.7: Model scores for the cluster in Figure 4.2a. Each Figure shows how well one of the submodels of PARma matches when aligned to the 7-mer that starts at the corresponding position. For instance in Figure 4.7a, the maximal value belongs to the 7-mer *TGCTGCT* and indicates that all observed and not observed T to C conversions match very well, when *TGCTGCT* is the microRNA seed site. A microRNA targeting the seed site *CACATTG* (corresponding to the secondary peak upstream of *TGCTGCT*) is also likely to lead to the observed conversion. The cleavage scores in Figures 4.7b and 4.7c indicate how likely the observed RNase T1 cleavages are, given the seed site is at the corresponding position. Both submodels would allow seed sites to start within a small window of about 10 bases and indicate that the secondary peak from 4.7a is unlikely to correspond to the true microRNA seed site. However, they agree with the primary peak of the conversion scores. Finally, the k-mer activity scores in Figure 4.7d indicate how many other PAR-CLIP clusters are likely to be explained by the corresponding k-mer and also points to the 7-mer *TGCTGCT*. This is indeed the 7mer-m8 seed site for miR-15a, which has been experimentally validated to target this cluster [Kiriakidou et al., 2004].

encodes 25 mature microRNAs, some of which are highly expressed in BCBL1 [Dölken et al., 2010]. Thus, PAR-CLIP clusters that are assigned to one of the KSHV microRNAs must not be present in DG75 and we can use the number of KSHV assigned PAR-CLIP clusters in DG75 as a measure of assignment accuracy. Although both cell lines, BC3 and BC1 in the PAR-CLIP data from Gottwein et al. [2011] are infected by KSHV, only BC1 is coinfectd by Epstein-Barr-Virus (EBV), which encodes 44 mature microRNAs. Hence, PAR-CLIP clusters that are assigned to one of the EBV microRNAs must not be present in BC1.

With respect to exclusive sites, PARma is more accurate than all other methods including PARalyzer independent of the dataset used for evaluation (see Figures 4.8a and 4.8d). More than 70% of all clusters, where PARma assigned a KSHV or EBV microRNA, only have reads in BCBL1 or BC1, respectively. This number drops to about 50%, when any seed match of a KSHV microRNA in a cluster is taken as evidence for a KSHV target site (*all.cluster*) or PARalyzer is used. When a seed match immediatly downstream of the main cross-linking site is used (*first.xlink*), the accuracy is almost as high as for PARma, but is heavily dependent on both dataset and the number of microRNAs used. Additionally, PARma's accuracy is significantly higher when it is run starting with all 16,384 7-mers (PARma) instead of microRNA 7-mer seeds only (PARma_miR). This suggests, that in several cases, there are seeds of KSHV/EBV

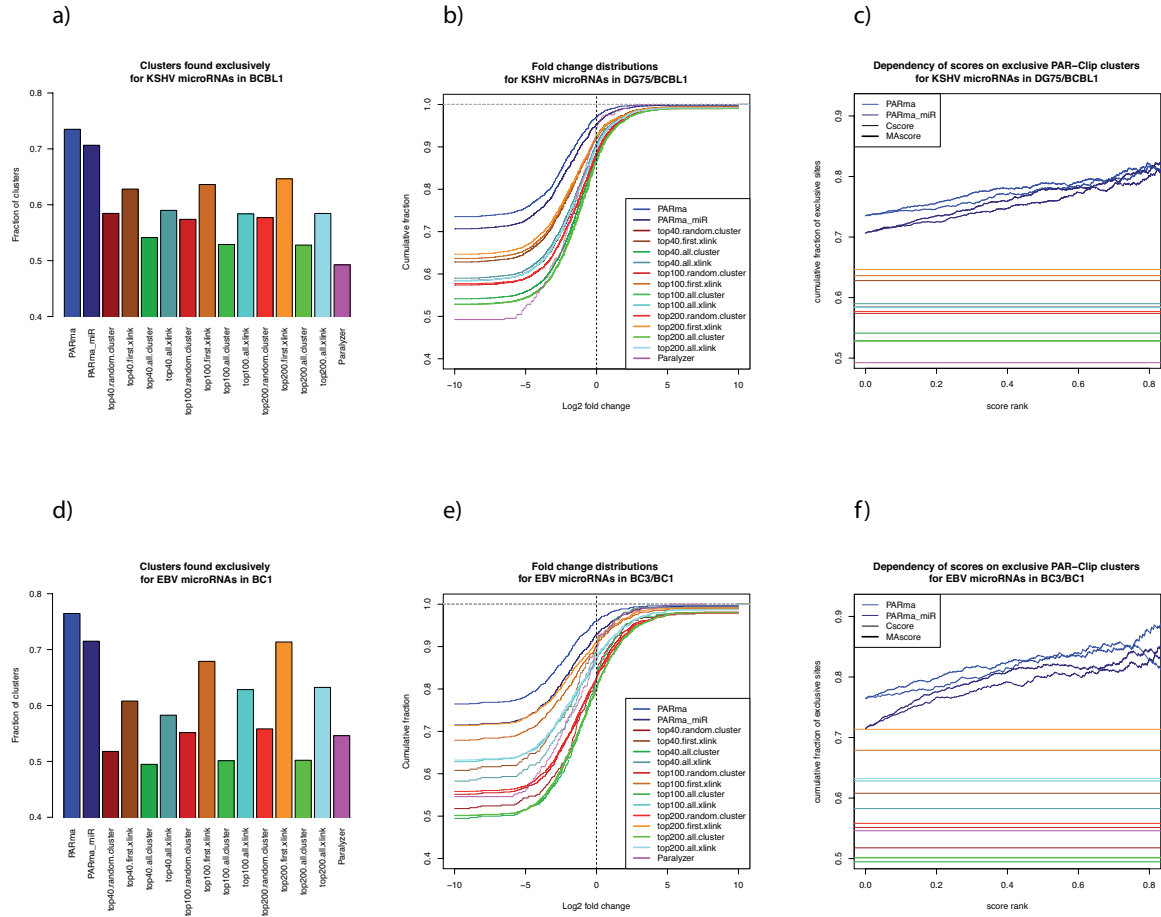


Figure 4.8: Evaluation using differential PAR-CLIP. KSHV microRNA target sites should only have reads in KSHV infected cell lines (Figures 4.8a-c) and EBV microRNA target sites should be exclusive to EBV infected cell lines (Figures 4.8d-f). PARma assigned KSHV microRNA target sites show a higher fraction of exclusive sites than all other methods (Figures 4.8a, 4.8d; see main text for a description of all other methods) and PARma run without constraining to known microRNA seeds yielded a higher fraction of exclusive sites than PARma using seeds as prior. Figures 4.8b and 4.8e show the log fold changes (control/infected) of PAR-CLIP read counts for clusters assigned to KSHV and EBV microRNAs, respectively. The log fold change of exclusive clusters (i.e. clusters that have no reads in one of the experiments) has been set to -10 or 10. PARma does not only have the largest fraction of exclusive clusters in both datasets (compare the left ends of Figures 4.8b and 4.8e to Figures 4.8a and 4.8d, respectively) but also the smallest fraction of KSHV or EBV clusters that have more reads in the KSHV or EBV negative cell line. The dependency of scores on the accuracy is shown in Figures 4.8c and 4.8f. In both datasets and for both scores, accuracy increases when more and more low scoring clusters are removed. As a reference, the accuracies of the other assignment methods are indicated with the same colors as in Figures 4.8b and 4.8e.

microRNAs in a non-exclusive cluster but there are also other overrepresented 7-mers that explain the conversions and RNase cleavage sites better.

We noticed that often, random reads are scattered across expressed transcripts in all experiments. Thus, a true KSHV microRNA target site may get random reads in the KSHV negative cell line (DG75) and therefore, may not be exclusively present in BCBL1. Therefore, we considered the number of PAR-CLIP reads in each KSHV or EBV microRNA assigned cluster and plotted their log fold change of DG75/BCBL1 or BC3/BC1, respectively (see Figures 4.8b and 4.8e). Independent on the fold change cutoff, PARma consistently identifies more KSHV or EBV microRNA clusters that have less reads in DG75 than in BCBL1 or in BC3 than in BC1, respectively. Specifically, less than 5% of KSHV cluster have more reads in DG75 than in BCBL1 for PARma assignments, which drops to below 90% for the other assignments.

In order to evaluate the computed Cscores and MAScores (see Methods section), we sorted clusters according to Cscore or MAScore and computed the fraction of BCBL1 and BC1 exclusive sites for KSHV and EBV microRNA assigned clusters, respectively. For both datasets the accuracy increases, when more and more of the low scoring clusters or clusters with multiple possible microRNAs are removed, achieving accuracies of 80% or more (see Figures 4.8c and 4.8f).

4.3.8 Validation against RIP-Chip data

To further validate target sites and target site assignments that are only found by PARma, and to invalidate target sites that have not been detected by PARma but by other methods, we considered RIP-Chip data that we measured for the cell lines DG75 and BCBL1 [Dölken et al., 2010]. In a RIP-Chip experiment, the amount of an RNA co-immunoprecipitated using an anti-AGO2 antibody is compared to RNA from a control-IP using microarrays. Thus, it measures the recruitment of an mRNA to Ago2-complexes in a quantitative way and is an alternative technique to PAR-CLIP to determine microRNA targets. Using proper data analysis methods [Erhard et al., 2013b], the differential enrichment of mRNAs with RISC can be computed between BCBL1 and DG75, which indicates, whether an mRNA is stronger associated with RISC in BCBL1 than in DG75. On average, this must be the case for targets of KSHV microRNAs.

Thus, we determined all genes that contain a KSHV microRNA target site according to PARma and Paralyzer (*both*), that contain a KSHV microRNA target site according to PARma and no KSHV microRNA target site according to Paralyzer (*PARma only*) and that contain a KSHV microRNA target site according to Paralyzer only (*Paralyzer only*) and compared it to genes without KSHV microRNA target sites (*none*; see Figure 4.9a). The *both* and *PARma only* genes showed significantly elevated differential RIP-Chip enrichment values ($p < 2 \times 10^{-4}$ and $p < 2 \times 10^{-7}$, respectively, one-sided Kolmogorov-Smirnov test), whereas *Paralyzer only* and *none* genes were indistinguishable from background. Thus, based on the RIP-Chip data, PARma effectively gets rid of false positive target sites detected by Paralyzer, and, in addition, picks up false negatives not detected by Paralyzer. We also repeated the same analysis for other methods replacing Paralyzer with similar results (see Figure 4.9b).

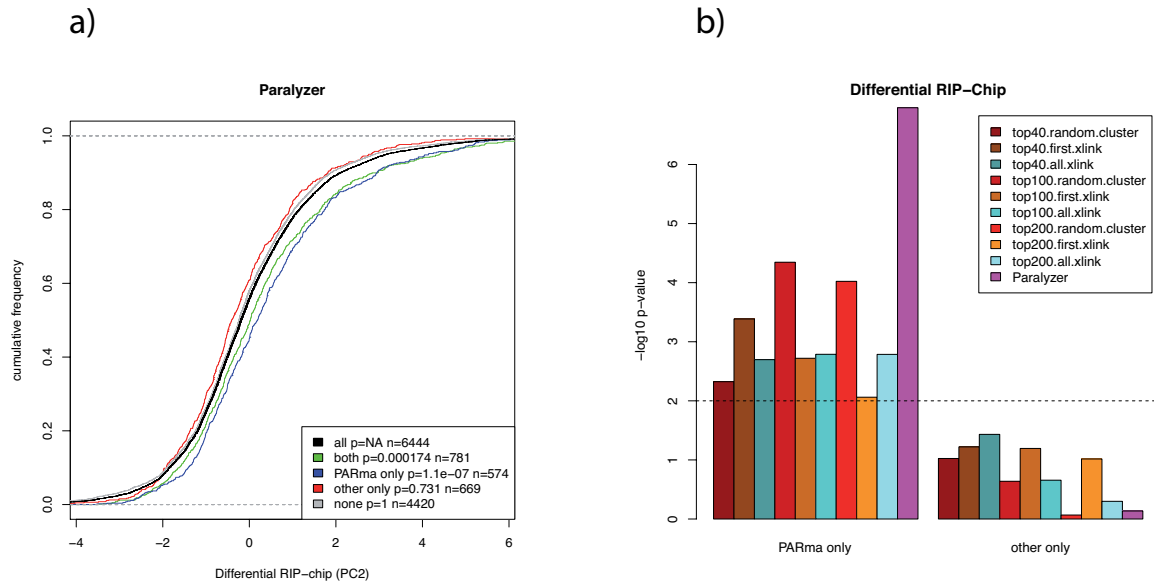


Figure 4.9: Validation against RIP-Chip. Figure 4.9a shows the distribution of differential RIP-Chip enrichments (PC2 scores) of BCBL1 and DG75 for different sets of PAR-CLIP targets. Higher values indicate a stronger enrichment of a gene with RISC in BCBL1 than in DG75, and, therefore, a set of KSHV microRNA targets should have a right-shifted distribution of PC2 scores. Genes that have been identified by PARma as well as Paralyzer to be KSHV microRNA targets indeed show such a shift, as well as genes that have only been found by PARma and not by Paralyzer (PARma only). In contrast, genes that are not targets of KSHV microRNAs according to both PARma and Paralyzer do not show a shift. Interestingly, genes found exclusively by Paralyzer and not by PARma are not shifted as well. We repeated this analysis for all other methods, as illustrated in Figure 4.9b. The p-values for the comparisons of PARma to all other methods indicate that PARma not only outperforms Paralyzer but all other methods as well.

4.4 Discussion

4.4.1 PAR-CLIP clusters

In this paper, we present a in-depth investigation of seed sites in PAR-CLIP clusters. The standard approach to assign microRNAs in all so far published PAR-CLIP studies [Hafner et al., 2010; Gottwein et al., 2011; Lipchina et al., 2011; Skalsky et al., 2012] was to select the top N expressed microRNAs and identify seed sites in the respective PAR-CLIP clusters. However, it is not clear, how N must be chosen: For small N, only a small fraction of clusters can be assigned and for larger N, cluster assignments get more and more ambiguous. Furthermore, independent on the choice of N or the exact way of searching for seeds, microRNA expression correlates only poorly with the number of clusters. Also, there are multiple studies reporting small RNAs other than microRNAs that are associated with the RISC. Thus, it seems advantageous to remove the restriction of searching for a predefined set of seeds.

PARma can be used for both searching for a predefined set of seeds and for an unconstrained search for all possible k-mers. In both cases, the assigned seeds fulfill two conditions in each cluster: First, the observed T to C conversions and RNase cleavage sites relative to the seed position match a model derived from all clusters and second, the seed site sequence is overrepresented. As illustrated in Figure 4.8, the unrestricted search is even more accurate in terms of assigning KSHV or EBV microRNAs to clusters that are exclusively present in KSHV or EBV infected cells, respectively.

We propose that the general approach of PARma can also be applied to other kinds of CLIP data. For instance, for iCLIP data [König et al., 2010], reads in valid target sites should start immediately after cross-linking sites. These specific start positions could be incorporated into an iCLIP model instead of the PAR-CLIP model of conversions and RNase T1 cleavage sites. However, how effective it is to exploit these characteristics of iCLIP data remains to be seen when more and more iCLIP data becomes available.

Clusters from a CLIP experiment are not necessarily true binding sites of the protein of interest: Neither the immunoprecipitation (IP) step nor the gel separation are 100% specific and thus, there may be artefacts of other RNA binding proteins (RBPs). If 40 distinct microRNA seeds are considered and matched to such clusters, more than 20% of the unspecific clusters are expected to contain at least one seed match by chance (assuming an average cluster length of 30 bp and a seed length of 6). This increases to almost 70%, when 200 microRNA seeds are considered. Thus, we expect that there is a considerable amount of false positive microRNA target sites in current PAR-CLIP datasets. Finding a reliable way of scoring clusters in order to filter such false positives is therefore of great importance.

To our knowledge, PARma is the first method to provide a scoring system that has been proven to improve accuracy upon filtering. The rationale for that is that there is no reason why unspecific clusters should match our PAR-CLIP model. Indeed, Cscores of intronic clusters, which likely are the result of unspecific IPs of other RBPs, are significantly lower than Cscores of 3'-UTR clusters (data not shown) in both AGO-PAR-CLIP datasets, which is in agreement with known mechanisms of microRNAs. Furthermore, even if unspecific clusters may match the PAR-CLIP model by chance and contain active k-mers by chance, it is unlikely that these k-mers occur at a position that matches the model. Thus, both Cscore and MAscore are expected to improve accuracy (see also Figure 4.8c and 4.8f).

4.4.2 PARma

For the conversion model used in PARma, we assume that cross-linking events are independent from each other. This means that given an uridine at a certain position relative to the seed site, the probability that a cross-linking event takes place and is sequenced at this position is not dependent on the location of other uridines. This assumption may be wrong, if one of the other uridines is already cross-linked. However, the probability that two cross-links can occur in close vicinity to each other is very low, since the incorporation rate of 4-thiouridine (4sU) is only about 1/40 and only 4sU gets cross-linked with high frequency at the wavelength used in PAR-CLIP [Hafner et al., 2010]. In addition, the reverse transcriptase (RT) is known to be rather inefficient in reading through the peptide chain still cross-linked to the 4sU-residue (which is responsible

for the U to C transition). Therefore, it becomes rather unlikely that the RT reads through two cross-links in a single RNA fragment.

Note also, that the model for conversions is not simply built by summing all cross-linking events for each position globally over all clusters. Such a procedure would be heavily influenced by a few clusters that have thousands of reads in comparison to the many clusters having only a few dozen reads. In contrast, our parameter estimation for the conversion model does not only exploit all clusters, but is also robust against outlier clusters by using robust regression and quadratic programming. Robustness in the parameter estimation is an important issue especially in the initial iterations. This is because seeds are not yet assigned with high confidence leading to many outliers.

PARma does not necessarily assign seed sites directly downstream of cross-linking sites. When the next uridine upstream of a true seed site is several nucleotides away, it may still get cross-linked. In such a case, PARma may still find another k -mer closer to the cross-linking site, dependent on the sequences, on other cross-linking events in the same cluster and on the RNase cleavage sites. However, PARma will report a low MAScore, since the other position will score similarly well.

PARma can be run for different values of k . The smallest reliable seed used in the literature is microRNA bases 2-7 [Bartel, 2009; Kertesz et al., 2007; Grimson et al., 2007; Friedman et al., 2008; Krek et al., 2005]. However, we noticed that PARma with $k = 6$ resulted in slightly worse accuracies for both our data sets in comparison to $k = 7$ (data not shown). This may be a consequence from the fact that random 6-mers are expected to occur every 4096 bases, and thus, every ~ 100 clusters (median length of clusters is 47). When at least 100 microRNAs with different 6-mer seeds are considered, every single cluster would on average have a seed match by chance. Thus, kmerExplain may have difficulties to reliably extract the signal of overrepresented 6-mers.

By the requirement that only a single k -mer is enough to explain a cluster, kmerExplain is able to avoid *overrepresented partial k-mers*: Consider the 7-mer-A1 seed site UCGUCGA that is explaining hundreds of clusters. Obviously, the sequence CGUCGAG is expected to be present in $\frac{1}{4}$ of these clusters and is thus highly overrepresented in the collection of all clusters. This overrepresented partial k -mer may also occur in additional clusters, i.e. without the leading U. Even if it is not overrepresented by itself but only due to an overlapping k -mer that is truly overrepresented, all additional occurrences may be mistaken for the seed site of a targeting microRNA not because the microRNA is active but only because of the overlap to an active microRNA seed. Obviously, kmerExplain avoids such overrepresented partial k -mers by the requirement that only a single k -mer can explain a cluster.

4.4.3 Comparison to PARalyzer

PARalyzer is a software package specifically designed for the analysis of PAR-CLIP data [Corcoran et al., 2011]. It utilizes kernel density estimation to estimate the probability of interaction along each cluster based on the normalized numbers of conversions and non-conversions at each position. There are two main differences to the basic approach from Hafner et al. [2010]: First, an interaction site is called when the estimated density of conversions is

greater than the estimated density of non-conversions instead of using the main cross-linking site for all clusters, which are filtered by certain criteria, and second, due to the kernel, the neighborhood of uridine sites is incorporated dependent on an arbitrarily chosen bandwidth parameter. It is unclear whether this approach is able to effectively filter out unspecific clusters. In addition, the pattern discovery module does not incorporate the information of cross-linking or RNase cleavage positions and is, thus, unable to resolve and score ambiguous seed matches. Furthermore, the PARalyzer pipeline does not include methods to handle spliced reads and, therefore, all studies that have used PARalyzer [Corcoran et al., 2011; Gottwein et al., 2011; Skalsky et al., 2012] may have missed all target sites that span exon-exon junctions. In the datasets we analyzed, 22.4% of all clusters in the coding region of transcripts span splice junctions (about 6% of all clusters).

4.4.4 Differential PAR-CLIP

In order to evaluate PARma, we directly compared the number of PAR-CLIP sequencing reads from multiple experiments mapped to each individual cluster. Our evaluation is based on the following consideration: When a cluster represents a valid target site of a KSHV microRNA, for instance, AGO should not be associated with it in KSHV negative cells and, therefore, the corresponding PAR-CLIP experiment should not yield sequencing reads mapping to this cluster (exclusive clusters).

While this is true for $\sim 80\%$ of all clusters assigned to a KSHV or EBV microRNA in both of the respective datasets when PARma is used (see Figures 4.8b and 4.8e), there is a considerable number of clusters, where this is not true. There may be several reasons for these: First, there is a considerable amount of background in the data, i.e. sequencing reads that are not due to specific cross-linking to AGO and indeed, almost all clusters have a positive log2 fold change of PAR-CLIP reads, which may be a consequence of background. Second, a target site could be targeted by multiple microRNAs. This is very probable for seed homologous viral microRNAs (e.g. kshv-miR-K12-11 has the same seed as hsa-miR-155), but may also come from strongly overlapping target sites. Accuracy increases when clusters are filtered by MAScore (see Figure 4.8c and 4.8f), which also indicates ambiguous assignments. Third, clusters may not be valid target sites and just by chance contain seeds of KSHV or EBV microRNAs, respectively, since accuracy also increases when clusters are filtered by Cscore.

It would be of great benefit to be able to convert our scores to a false discovery rate as a statistically meaningful measure. This could be done if there was a way to determine how many of the non exclusive clusters are still valid KSHV or EBV target sites. However, it is difficult to estimate the background, which is dependent on transcript expression, on other RNA binding proteins that target these transcripts and probably on many more factors. Additionally, the extent of overlapping or truly ambiguous target sites is unclear. Furthermore, the presence of reads is subject to stochastic sampling effects due to the relatively small numbers of reads. Thus, it is currently not possible to estimate reliable false discovery rates based on differential PAR-CLIP.

Conclusion

In this paper we presented PARma, a method to analyze PAR-CLIP data. Clusters are defined in a similar way as before [Hafner et al., 2010; Corcoran et al., 2011]. The main purpose of PARma is a) to define reliable microRNA target sites and b) to identify the microRNA responsible for each identified target site. Therefore, two scores are computed: The Cscore assesses the likelihood that a cluster is a valid microRNA target site and the MAScore corresponds to the confidence that the assigned microRNA is the true regulator.

PARma utilizes features specific to PAR-CLIP data to determine seed sites: The positions of cross-linking sites and missing cross-linkings as well as cleavage sites of RNase T1 relative to seed sites are learned and incorporated into a generative model. This model is used to guide a novel pattern discovery tool, kmerExplain, that estimates activity probabilities for k-mers.

Our method can be used to discover active k-mers in an unbiased manner, i.e. without assuming a set of admissible k-mers such as the top N microRNA seeds. Each reported active k-mer nevertheless has two properties: It explains several clusters and the positions where it occurs match the model of PAR-CLIP data learned from all target sites.

Using differential PAR-CLIP data, we have shown that PARma is more accurate than existing methods and that both Cscore and MAScore are useful measures to rank clusters.

4.5 Methods

4.5.1 Data

The data from Gottwein et al. [2011] has been downloaded from GEO (accession number: GSE32113). DG75 and BCBL1 PAR-CLIP experiments have been performed as described [Kishore et al., 2011; Jaskiewicz et al., 2012]. Briefly, a total of $3 * 10^8$ cells per replicate were grown and treated with 4-thiouridine (Sigma) for 14 hours (final concentration $100 \mu\text{M}$). Cells were pelleted and washed in cold PBS. Aliquotes of $5 * 10^7$ cells were resuspended in 5 ml of cold PBS, placed in a 15 cm petri dish and irradiated at 365 nm with 100 mJ twice on ice, with 30 s break in between. Crosslinked cells were collected, pelleted and snap-frozen. PAR-CLIP was performed using 11A9 anti-Ago2 monoclonal antibody [Rüdel et al., 2008]. PAR-CLIP sequencing data have been deposited at GEO (accession number: GSE43909).

4.5.2 Raw data processing and cluster definition

The deep sequencing data have been processed using an in-house pipeline consisting of adapter trimming, read mapping against genomes and transcriptomes, integrating all mappings and cluster identification as well as filtering.

Read mapping

The 3' sequencing adapter sequence are trimmed from each sequencing read using a specially tailored sequence alignment variant that aligns a prefix of the adapter sequence to a suffix of each sequencing read. After that, equal sequences are collapsed and mapped to the human genome (hg19), the KSHV genome (NC_009333.1), the EBV genome (NC_009334.1) and the human transcriptome (Ensembl v60) using Bowtie version 0.12.7 [Langmead et al., 2009]. For each collapsed read, all mappings for an experiment are then collected and the best in terms of mismatches is written to a single BED file for each experiment including information about the read count (number of sequences before collapsing), the mismatches of each alignment and the number of alignments after mapping transcriptome alignments to the genome. Here, T to C conversions are not counted as mismatches, since they are expected due to the experimental protocol.

Cluster identification

All BED files are then simultaneously scanned chromosome by chromosome in a strand specific manner and overlapping reads are clustered. We use only reads without mismatches (except for T to C conversions). Clusters are then filtered according to similar criteria as before [Hafner et al., 2010; Corcoran et al., 2011]: read count at least 5, at least 3 distinct read species. Clusters are quantified using the count of the main cross-linking site. After clustering, normalization factors are computed such that the median fold change to a reference experiment (we took the one with the most reads) is 1. Then, in a second pass, all clusters are removed where all experiments have less than 10 normalized read counts.

We also implemented three additional options: First, it is known that two target sites may overlap. Especially for viral microRNAs, several of such cases are known [Nachmani et al., 2010]. Thus, we split each cluster: Only reads spanning the main cross-linking site are used and the criteria from above are checked. Then, the main cross-linking site of the remaining reads is determined. This is repeated as long as all criteria are fulfilled.

Second, since target sites may span splice junctions and we mapped reads to the transcriptome, we can also identify spliced PAR-CLIP clusters. However, when allowing for spliced reads, the definition of a cluster is not straight-forward: For instance, for a 3' end of an exon, there may be reads starting in the exon and ending in the neighboring intron and reads that connect this exon to various other exons. We resolve such inconsistencies by first removing all exon-intron reads and then by removing reads to exons with fewer reads, if necessary.

Third, since target sites may be wider than the maximal sequence length, we extend all untrimmed reads up to the next RNase T1 cleavage site (i.e. after the next G). This is important because in the following, we specifically use these cleavage sites in our generative model.

Visualization

In order to visualize PAR-CLIP data appropriately, we developed a specialized web-based visualization tool (see Figure 4.2). Other than the widely used genome browsers from UCSC or Ensembl, our viewer offers specialized visualization tools for PAR-CLIP data: We visualize

several evolutionary conservation scores, including k-mer branch lengths that have been used for microRNA target prediction [Friedman et al., 2008], sequence read coverage, SNPs, the actual reads with indicated conversions, conversion densities, transcripts and PAR-CLIP clusters. Other than genome browsers, our viewer is able to shrink introns in a data dependent way (i.e. if there are no reads mapped to an intron, it is not visualized in scale to the exons but shrunk to a few pixels). This is a major advantage to showing everything in scale when visualizing transcript related data, since usually the long introns are often not of interest in contrast to the short exons.

4.5.3 PARma

The result of our preprocessing, which is very similar to previous work [Hafner et al., 2010; Corcoran et al., 2011], is a set of clusters \mathbb{L} . Each cluster $L \in \mathbb{L}$ is characterized by its sequence $s(L)$, its conversion profile $conv_L$ and two vectors $start_L$ and end_L . $conv_L$ is a vector containing for each position within L the number of conversions, whereas $start_L$ and end_L contain for each position the number of reads starting and ending there, respectively. Furthermore, we define $T(L) = \{i \in \{1..|s(L)|\} | s(L)_i = T\}$ as the set of possible conversion sites and $G(L) = \{i \in \{1..|s(L)|\} | s(L)_i = G\}$ as the set of possible RNase T1 cleavage sites.

Model fitting

The PARma model consists of three submodels, incorporating T to C conversion data, 5' RNase cleavage data and 3' RNase cleavage data, respectively. The conversion model assigns each position i relative to the seed site a cross-linking probability $xlink(i)$. Then, the cross-linking score s_{xlink} for a seed position j in cluster L can be computed as

$$s_{xlink}(L, j) = \frac{\sum_{k \in T(L)} conv_L(k) \cdot xlink(j - k)}{\sum_{k \in T(L)} conv_L(k) \cdot \sum_{k \in T(L)} xlink(j - k)}$$

This is essentially the normalized dot product of two vectors: The first vector contains the observed conversion counts for all conversion positions, the second contains the cross-linking probabilities for these positions. Thus, $s_{xlink}(L, j) = 1$ if and only if the observed conversions exactly meet the expected cross-links and approaches 0 when the observed counts differ from the expected. Note that $xlink$ must only be known up to a constant factor. This allows us to fit the model without making any further assumptions: Given a current estimate j of the seed position for each cluster L , we first estimate the ratio $R_{k,l}$ for each pair of model positions k and l by collecting all clusters L with $j - k \in T(L)$ and $j - l \in T(L)$. Then we use robust linear regression to fit a line through the origin given the values $conv_L(j - l)$ and $conv_L(j - k)$ of all collected clusters L . The slope of this line then is a robust estimate of $R_{k,l}$. Given the estimates of $R_{k,l}$ for all $k < l$, we obtain the final estimate of $xlink$ by minimizing

$$\sum_{k,l} \left(\frac{xlink(k)}{xlink(l)} - R_{k,l} \right)^2$$

subject to $xlink(j) \geq 0$ and $\sum_j xlink(j) = 1$ using quadratic programming. Note that the final constraint arbitrarily fixes the above mentioned constant factor and is necessary to get a unique solution.

The 3' RNase cleavage model assigns each position i relative to the seed site the cumulative probability $c3(i)$, that the RNase cleavage site is $\leq i$. Given a cluster L , let $G(L) = \{k_1, \dots, k_n\}$ with $k_{i-1} < k_i$. Then, the downstream cleavage score $s_{downstream}$ for a seed position j in cluster L can be computed as

$$s_{downstream}(L, j) = \frac{\sum_{i \in 1..n} end_L(k_i) \cdot p(k_i)}{\sum_{i \in 1..n} end_L(k_i)}$$

$$p(k_0) = c3(j - k_0)$$

$$p(k_i) = c3(j - k_i) - c3(j - k_{i-1})$$

Note that we use cumulative probabilities here: In contrast to cross-linking positions, RNase cleavage sites are not independent: For instance, let cluster L_1 have two consecutive G 5 bp downstream of the true seed site ($=SEED=NNNNNGG...$) and cluster L_2 only one G 6 bp downstream of its true seed site ($=SEED=NNNNNNG...$). The second G in L_1 is at the same position relative to the seed site as the single G in L_2 . The RNase may have enough room to cut after the first G in L_1 and thus, all reads in L_1 may end 5 bp downstream of the seed site. In cluster L_2 , all reads will end 6 bp downstream of the seed site. Thus, depending on where other G sites are located, read end probabilities will differ. Using cumulative probabilities in the model and computing the probabilities depending on G locations from cumulative probabilities is able to alleviate this problem. $c3$ is estimated by using the current estimates j of the seed position for each cluster L . The cumulative probability then is the number of times a position is upstream of the main RNase cleavage site divided by the number of clusters.

The 5' RNase cleavage model is formulated analogously to the 3' model. The final score for a position j in cluster L_i then is calculated as the product of the three submodel scores:

$$p_{i,j} = s_{xlink}(L_i, j) \cdot s_{downstream}(L_i, j) \cdot s_{upstream}(L_i, j)$$

KmerExplain

Given a set of sequences $\mathbb{S} = \{S_1, \dots, S_n\}$ and scores $p_{i,j}$ for each position j in cluster L_i , kmerExplain estimates k-mer activity probabilities using an EM algorithm for the following probabilistic model: We assume that each sequence is generated by only a single k-mer. Then, the probability of generating a sequence S by a k-mer at its j th position is

$$P(S|j) = \alpha_{S^j} \cdot \prod_{c \neq j} (1 - \alpha_{S^c})$$

Here, α_x is the activity probability of k-mer x and S^j denotes the j th k-mer in S . The likelihood of \mathbb{S} then is

$$P(\mathbb{S}) = \prod_{i=1}^n P(S_i) = \prod_{i=1}^n \sum_j P(S_i|j) p_{i,j}$$

Thus, we have to estimate α_x for all k-mers x under hidden parameters j (active k-mer position in S_i). In the E-step we compute the values $q_{i,j}$ given the current estimates of α_x as

$$q_{i,j} = \frac{p_{i,j} P(S_i|j)}{\sum_c p_{i,c} P(S_i|c)}$$

The values $q_{i,j}$ represent current estimates of the probability $P(j|S_i)$. The estimator for α_x then is (M-step; $\delta_{x=y}$ is the Kronecker delta: $\delta_{x=y} = 1 \Leftrightarrow x = y$):

$$\alpha_x = \frac{1}{n} \sum_{i,j} q_{i,j} \cdot \delta_{x=S_i^j} \quad (4.1)$$

Proof: The conditional expected value of the log likelihood and its partial derivative with respect to α_x are:

$$\mathbb{E} = \sum_{i,j} q_{i,j} \log P(S_i|j) \quad (4.2)$$

$$= \sum_{i,j} q_{i,j} \log \left(\alpha_{S_i^j} \cdot \prod_{c \neq j} (1 - \alpha_{S_i^c}) \right) \quad (4.3)$$

$$\frac{\delta \mathbb{E}}{\delta \alpha_x} = \frac{1}{\alpha_x} Q_x - \frac{1}{1 - \alpha_x} Q_{\bar{x}} \quad (4.4)$$

$$Q_x = \sum_{i,j} q_{i,j} \cdot \delta_{x=S_i^j} \quad (4.5)$$

$$Q_{\bar{x}} = \sum_{i,j} q_{i,j} \cdot (1 - \delta_{x=S_i^j}) \quad (4.6)$$

Respecting that $Q_x + Q_{\bar{x}} = n$, setting (4.4) to zero and solving for α_x yields equation (4.1). \square

Final assignment and integration

The output of the final iteration consists of scores $p_{i,j}$ for each position j in cluster L_i as well as $q_{i,j}$, which are estimates of the probability $P(j|S_i)$. The first is a quantity indicating how well the experimental data fits the model, when *any* k-mer at position j has generated cluster L_i . The latter also incorporates the k-mer activity probability (i.e. how well does the experimental data fit the model, when the given k-mer at position j has generated cluster L_i). Furthermore, for each cluster L_i we get the most probable k-mer generating this cluster at position $g_i = \operatorname{argmax}_j \{q_{i,j}\}$. We use these quantities to compute confidence scores for each cluster (Cscore) and each k-mer assignment (MAscore):

$$\text{Cscore}(i) = p_{i,g_i} \quad (4.7)$$

$$\text{MAcore}(i) = \frac{q_{i,g_i}}{\sum_j q_{i,j}} \quad (4.8)$$

We integrate multiple experiments (either replicates of the same condition or multiple conditions) by first running PARma for each experiment individually and then taking the generating k-mer by computing a weighted sum over all $q_{i,j}$ from all experiments (weighted by the respective read count in the cluster) and taking the maximum. The Cscore then is the weighted sum of the p_{i,g_i} values and the MAcore the maximal MAcore of all experiments at this position.

4.6 Software availability

PARma is published under the GNU General Public License v3 and at the project website <http://www.bio.ifi.lmu.de/PARma>.

Chapter 5

RIP-chip enrichment analysis

Motivation: In the previous chapter I presented a method to accurately identify microRNA targets in PAR-CLIP data. PAR-CLIP is a relatively new experimental technique to discover microRNA targets and involves several experimental steps that may potentially fail (see section 2.1.2). RIP-Chip is a more established method, it involves less experimental steps with the cost of only identifying target genes instead of target sites. Furthermore, it inherently involves a control experiment and, thus, RIP-Chip provides quantitative measurements of microRNA targets (i.e. how many copies of an mRNA are bound by RISC). In any case, RIP-Chip and PAR-CLIP both provide different aspects of microRNA targets and, therefore, complement each other. Thus, it is not only important to analyse PAR-CLIP data properly as described in the previous chapter, but also to handle RIP-Chip data in an appropriate way. Our collaboration partners generated several RIP-Chip datasets [Dölken et al., 2010], and based on observations made from these data, I developed analysis methods that address several issues associated with RIP-Chip data, as described in this chapter. Equally to PARma, these methods were also applied to the available datasets for the analyses presented in the next chapter.

Publication: This chapter has been published in Bioinformatics [Erhard et al., 2013b]. Here, I adapted the layout and made minor corrections to the text.

My contribution: I came up with the ideas and the methods, implemented the method, carried out evaluations and wrote the paper.

Contribution of co-authors: Lars Dölken provided RIP-Chip data and helped to revise the manuscript. Ralf Zimmer supervised the work and helped to revise the manuscript

5.1 Abstract

5.1.1 Motivation

RIP-chip is a high-throughput method to identify mRNAs that are targeted by RNA binding proteins. The protein of interest is immunoprecipitated and the identity and relative amount of mRNA associated with it is measured on microarrays. Even if a variety of methods is available to analyze microarray data, e.g. to detect differentially regulated genes, the additional experimental steps in RIP-chip require specialized methods. Here, we focus on two aspects of RIP-chip data: First, the efficiency of the immunoprecipitation step performed in the RIP-chip protocol varies in between different experiments introducing bias not existing in standard microarray experiments. This requires an additional normalization step to compare different samples and even technical replicates. Second, in contrast to standard differential gene expression experiments, the distribution of measurements is not normal. We exploit this fact to define a set of biologically relevant genes in a statistically meaningful way.

5.1.2 Results

Here, we propose two methods to analyse RIP-chip data: We model the measurement distribution as a gaussian mixture distribution, which allows us to compute false discovery rates (FDRs) for any cutoff. Thus, cutoffs can be chosen for any desired FDR. Furthermore, we use principal component analysis to determine the normalization factors necessary to remove immunoprecipitation bias. Both methods are evaluated on a large RIP-chip dataset measuring targets of Ago2, the major component of the microRNA guided RNA induced silencing complex (RISC). Using published HITS-CLIP experiments performed with the same cell line as used for RIP-chip, we show that the mixture modelling approach is a necessary step to remove background, that computed FDRs are valid and that the additional normalization is a necessary step to make experiments comparable.

5.1.3 Availability

An R implementation of REA is available on the project website (<http://www.bio.ifi.lmu.de/REA>) and as supplementary data file.

5.2 Introduction

Gene expression is a highly complex process that is controlled on multiple levels by various proteins and RNAs. Various experimental protocols have been established to measure expression levels of mRNAs or proteins, targets of transcription factors or post-transcriptional regulators and many other parameters of gene expression in a genome-wide manner. Each step of such a high-throughput experiment may introduce systematic errors (bias) or random variation (noise)

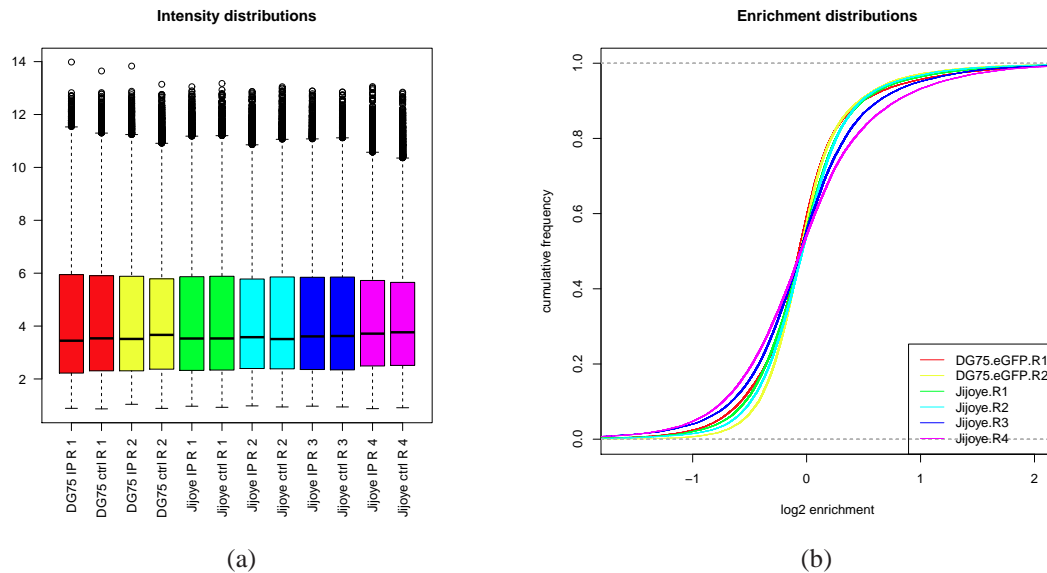


Figure 5.1: Measurement distributions for our Ago2 RIP-chip experiment. Boxplots for the intensity distributions of the measured microarrays described in the Methods section are shown in Figure 5.1a. Although the intensity distributions of the various arrays are properly normalized using RMA, the enrichment distributions are significantly different from each other (see Figure 5.1b). This is a consequence of differing IP efficiencies and must be accounted for when analyzing the respective data.

into the generated data and specialized methods are necessary to deal with particular kind of bias and noise and to answer specific questions using high-throughput data.

The most widely used high-throughput experiments are based on microarrays or next generation sequencing (NGS) and are designed to measure the amount of all mRNAs in one or multiple conditions [Malone and Oliver, 2011]. Based on the raw intensities from a microarray experiment or the sequencing reads from an NGS experiment, several analytical steps are taken, including normalization, summarization and statistical evaluation [Gentleman, 2005]. There is a vast amount of literature describing various methods fulfilling these steps to identify differentially regulated genes [Park et al., 2003; Fundel et al., 2008; Marioni et al., 2008; Wang et al., 2009; Irizarry et al., 2006].

Chromatin immunoprecipitation followed by microarray analysis or next generation sequencing (ChIP-chip/ChIP-seq) can determine the targets of DNA binding proteins and has successfully been applied to a wide range of transcription factors and cell types [Ren et al., 2000; Johnson et al., 2007; Birney et al., 2007]. In addition to the above mentioned analysis methods necessary for microarray and NGS data, it has been recognized that additional methods are necessary to successfully determine target sites on the genome and thus, a variety of methods is described in the literature [Zhu et al., 2010; Ho et al., 2011; Park, 2009].

In recent years it has become apparent that transcriptional regulation is only one part of the machinery carrying out gene regulation. RNA binding proteins (RBPs) and RNA binding ribonucleoproteins (RNPs) play important roles and are responsible for splicing, RNA editing, regulation of translation and RNA degradation [Witten and Ule, 2011; Nishikura, 2006; Bartel, 2009]. These processes are highly regulated by sequence-specific binding of RBPs or RNPs to the mRNA. MicroRNAs are small 20-24 nt long RNA molecules, that have emerged in recent years as important post-transcriptional regulators involved in all known multicellular organisms. They play important roles in development, tumorigenesis and viral infection. They act by guiding the RNA induced silencing complex (RISC) to mRNAs by binding to their 3'UTRs in a sequence specific manner, which leads to inhibition of translation or RNA degradation [Bartel, 2009].

A powerful experimental high-throughput technique to detect targets of RNA binding proteins or ribonucleoproteins such as RISC is based on immunoprecipitation (IP) of the RBP or RNP with associated mRNAs followed by microarray or NGS measurement (RIP-chip/RIP-seq) [Mukherjee et al., 2009; Hendrickson et al., 2008; Karginov et al., 2007; Stoecklin et al., 2008; Landthaler et al., 2008]. Targets of the RBP/RNP are enriched in the RIP experiment in comparison to a control measurement using an unspecific antibody or total RNA. Novel techniques including HITS-CLIP, iCLIP and PAR-CLIP also include crosslinking of the protein to the mRNA followed by digestion of the unprotected mRNA in order to determine the precise location of the target site [Chi et al., 2009; König et al., 2010; Hafner et al., 2010].

The main question in a RIP-chip experiment is to determine the set of target genes of the immunoprecipitated protein. A basic answer is a sorted list of *enrichment values* that can be computed for each gene by dividing the intensity value in the IP fraction microarray by the intensity in the control microarray. This is very similar to standard differential gene expression (DE) experiments: Here, differentially regulated genes can be determined by a sorted list of *fold changes* computed for each gene by dividing the intensity in condition A by the intensity in condition B. Consequently, RIP-chip data is often analyzed using standard methods borrowed from the DE setup such as fold changes [Stoecklin et al., 2008], t statistics [Mukherjee et al., 2009] or moderated t statistics [Hendrickson et al., 2008].

However, as indicated above, additional experimental steps may introduce additional bias: In contrast to log fold change distributions of DE experiments, log enrichment distributions of RIP-chip experiments are not normal but typically have heavier right tails (Mukherjee et al. [2009]; Dölken et al. [2010], see also Figure 5.1). This is an indication that RIP-chip is able to separate true targets from the background very efficiently. Here, we exploit these skewed distributions to estimate the biological significance of genes. Note that this is different from the statistical significance usually computed for DE experiments, where p-values are related to the reproducibility of the measurements and not to biological relevance.

The above mentioned question about the set of target genes in a RIP-chip experiment only considers a single condition, in contrast to a DE experiment. However, especially for RISC-IP experiments, an additional question is to determine differential microRNA targets between two or several conditions. For instance, if these conditions are *control* and *transfected microRNA* [Hendrickson et al., 2008], differential targets would be targets of the transfected microRNA, if there are *uninfected* and *virus infected* cells, differential targets would include targets of viral microRNAs [Dölken et al., 2010]. The answer to this question can be given by genes that are more

enriched in condition A than in condition B, either by choosing two cutoffs on the corresponding enrichment values (i.e. at least x fold enriched in A and at most y fold enriched in B) [Dölken et al., 2010], by computing *differential enrichment values* as the ratio of the two enrichment values [Hendrickson et al., 2008] or by a mixture of both approaches [Karginov et al., 2007].

All answers to this question necessarily have to compare enrichment values (i.e. ratios of intensities) of the conditions. However, IP efficiencies may vary between independent experiments, and it is important to account for this bias when comparing enrichment values. Obviously, the same problem exists for the summarization of replicate measurements (see Figures 5.1 and 5.6).

Here, we develop a suite of methods to properly analyze RIP-chip datasets, which take care of the unique properties of such data introduced by the IP: First, we use a gaussian mixture model approach to find statistically meaningful cutoffs for enrichment values. We show that this approach can be used to filter unexpressed genes, that it allows to compute false discovery rates (FDRs) for sets of biological significant genes and that it is in fact a necessary step to make experiments comparable to each other. We also address the problem of differing IP efficiencies by introducing a principal component analysis (PCA) based method to normalize enrichment distributions in a data dependent manner. We use publicly available HITS-CLIP data measured for the same cell lines [Riley et al., 2012a] as standard-of-truth for evaluation and show that the proposed methods provide significant improvements for the analysis of RIP-chip data.

5.3 Methods

5.3.1 Data processing

The RIP-chip data for this paper has been taken from our study of herpes viral RISC-IP experiments [Dölken et al., 2010]. Since the publications of the original study, additional replicates have been measured and all chips including the new ones have been processed as described [Dölken et al., 2010]. Briefly, RNA from Ago2-IPs and either BrdU-IPs or total RNA has been measured on Affymetrix GeneST arrays and all raw data have been normalized using RMA [Irizarry et al., 2003], \log_2 enrichment values have been computed by subtracting the control-IP/total RNA log intensity from the Ago2-IP log intensity for each probeset and each replicate experiment. Then, probesets have been mapped to Ensembl genes by using the annotation derived from Biomart. HITS-CLIP clusters (i.e. high confidence microRNA target sites) used to evaluate the mixture model approach for the cell line Jijoye has been downloaded from the supplementary data of Riley et al. [2012a]. We also repeated the same analysis using PAR-CLIP data for Jijoye that has been measured and analyzed as described in Hafner et al. [2010] in the lab of Markus Landthaler at the MDC Berlin (will be published elsewhere).

5.3.2 Mixture model fitting

Gaussian mixture models for sets of log enrichment values are fitted using the Mclust package in R [Fraley and Raftery, 2002]. Z-scores for each gene can then be computed using the background distribution:

$$zscore(g) = \frac{e(g) - \mu_{bg}}{\sigma_{bg}} \quad (5.1)$$

Here, $e(g)$ is the \log_2 enrichment of gene g , μ_{bg} and σ_{bg} are the mean and standard deviation of the background component of the gaussian mixture model (which we always take as the component with the smaller mean). The false discovery rate (FDR) for a cutoff c is defined as the expected fraction of background genes g_b with $e(g_b) > c$ under all genes g with $e(g) > c$:

$$FDR(c) = \frac{1 - cdf_{bg}(c)}{|\{g|e(g) > c\}|} \cdot |BG| \quad (5.2)$$

cdf_{bg} is the cumulative distribution function of the background component of the mixture model and $|BG|$ the expected number of background genes (estimated by the mixture model). This is mathematically equivalent the Benjamini-Hochberg multiple testing correction [Benjamini and Hochberg, 1995] for the onesided p-values derived from the z-scores in (5.1) multiplied with the expected fraction of background genes.

For the running window approach (see section 5.4.1), we first selected a window of w genes with the smallest Ago2-IP intensities and fitted the mixture model (first window). Then we removed the s genes with smallest Ago2-IP intensities and added the next s smallest still unselected genes and again fitted a mixture model. This step was repeated until the window reached the top Ago2-IP intensities. For the analyses we chose $w = 1000$ and $s = 20$.

We use two metrics to evaluate the fit of background and target distributions:

$$d(bg, t) = \frac{\mu_t - \mu_{bg}}{\sigma_{bg}} \quad (5.3)$$

$$skew(E) = -\log_{10}(ksp(E, -E)) \quad (5.4)$$

bg and t are the background and target components of the mixture model, respectively, E is the set of \log_2 enrichment values used to fit the mixture model and $ksp(E, -E)$ is the p-value of the Kolmogorov-Smirnov test comparing the distribution of E to the distribution of negated enrichment values $-E$. Thus, the distance score $d(bg, t)$ measures the distance of the background and target distributions with respect to the width of the background distribution, whereas the asymmetry score $skew(E)$ measures the skewness of the distribution without the need to fit a mixture model.

5.3.3 PCA

Principal component analysis is performed using the function `prcomp` in R. When there are k experiments/replicates and, therefore, k \log_2 enrichment values per gene, PCA is applied to the k -dimensional space of genes. The first principal component is the direction of the greatest variance, and is used to compute the summary enrichment value $\hat{e}(g)$ of gene g by taking the dot product of the replicate measurement z-scores $\langle z_1(g), \dots, z_k(g) \rangle$ and the direction of the first principal component $PC1$:

$$\hat{e}(g) = \langle z_1(g), \dots, z_k(g) \rangle \cdot PC1 \quad (5.5)$$

The geometrical interpretation of this weighted average of the replicate enrichment values is that the dot product does an orthogonal projection of the k dimensional point g onto the first principal component and measures the distance to the origin.

It is not necessary to center the points before PCA, since we perform PCA on the z-scores derived from the mixture modelling approach (see (5.1)) the point cloud is naturally centered at the means of the background distributions and centering at the overall mean may not be appropriate. Also, we perform PCA only on targets (as defined by an FDR of 1%). This is necessary, because if the number of background genes is much higher than the number of target genes, stochasticity in the background could mask the effects in the target genes to some extent.

Differential targets will deviate from this vector in a specific direction: E.g., if we have two replicates of two conditions A and B and, therefore, an enrichment vector $\langle z_{a_1}, z_{a_2}, z_{b_1}, z_{b_2} \rangle$, any gene that is target specifically in A has greater enrichment in A than in B: z_{a_1} and z_{a_2} is greater than z_{b_1} and z_{b_2} . Thus, if there are enough differential targets, the second principal component will point into the direction of the deviations of their enrichment vectors. Therefore, the summarized differential enrichment value $\hat{e}_d(g)$ can be computed similarly to the overall summary enrichment value in equation 5.5 by taking the dot product of the z-score vector $\langle z_1(g), \dots, z_k(g) \rangle$ and the direction of the second principal component $PC2$:

$$\hat{e}_d(g) = \langle z_1(g), \dots, z_k(g) \rangle \cdot PC2 \quad (5.6)$$

Note that both the enrichment value $\hat{e}(g)$ and the differential enrichment value $\hat{e}_d(g)$ incorporate a linear normalization that removes bias due to differing IP efficiencies.

5.4 Results

5.4.1 Select relevant genes

The first step in our analysis of RIP-chip data is the filtering of unexpressed genes. On modern microarrays such as the Affymetrix GeneST arrays used for our data, probesets against all known human genes are available. Even if virtually all probesets have non-zero intensities, we can expect that only a fraction of all genes is expressed in a specific condition. We noticed that the asymmetry of the log enrichment distribution is not observable over the whole range of IP intensities (see Figures 5.2a and S2). For low intensity genes, the distribution indeed looks normal, which is expected for a set of genes that is not or almost not expressed. Therefore, we employed a running window approach for fitting the mixture model (see Methods) and evaluated each window with respect to the distance of the two components of the fitted model and the extent of asymmetry (see Figure 4.7). At intensity values ≈ 5 a significant increase in both scores was observable. We chose to use all genes above an intensity cutoff where $d(bg, t) > 1$ and $skew(E) > 2$, i.e. where the means of the two mixture components are at least one standard deviation away from each other and where the asymmetry becomes significant with p-value 0.01.

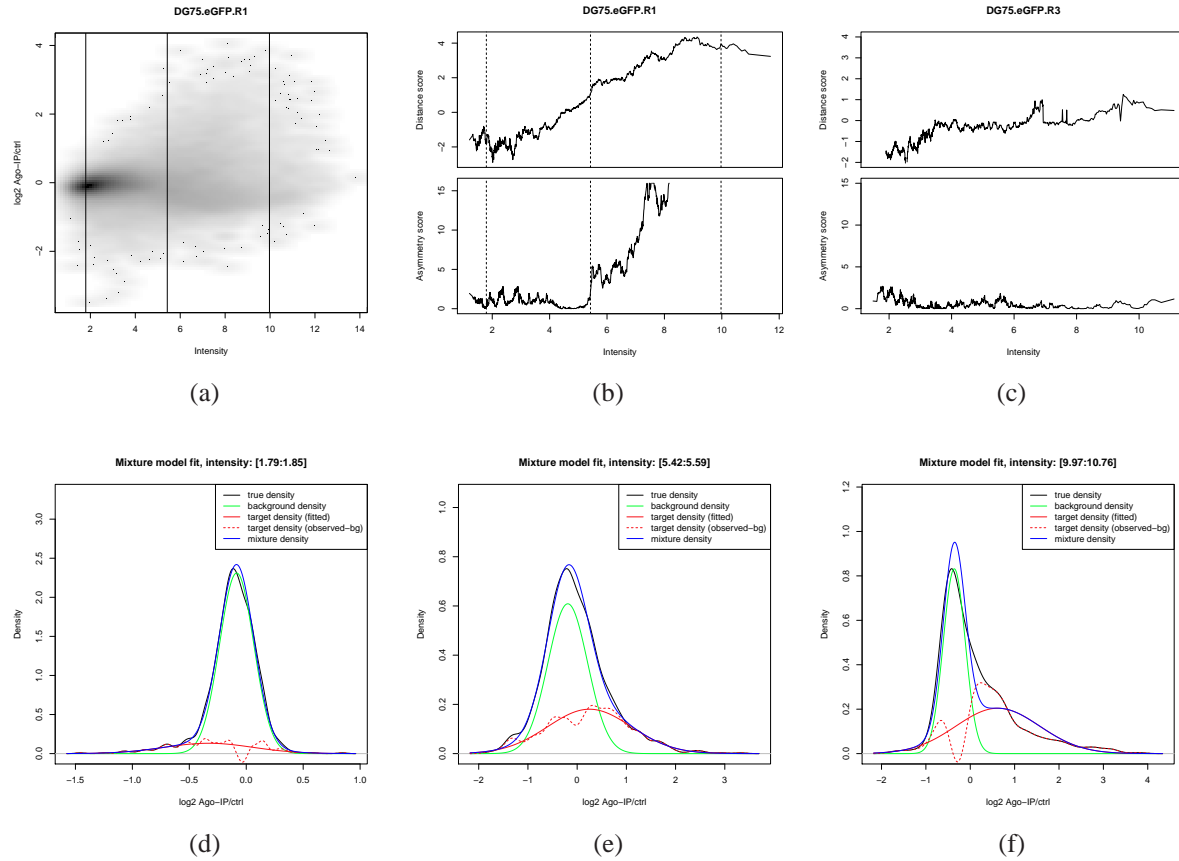


Figure 5.2: Selecting expressed genes. Figure 5.2a shows a density scatterplot of the \log_2 IP intensities against the \log_2 enrichment values of all genes for the first Jijoye RIP-chip replicate. In Figures 5.2b and 5.2c, the distance and asymmetry scores are plotted for all windows for replicate one and three of the DG75-eGFP experiment. Starting from intensity values of ≈ 5 , the distribution seems to be a mixture of two normal distributions. In contrast to the first replicate, the third does not show the expected behavior of the mixture of a background and target distribution which was a consequence of poor RNA quality in this experiment. The running window mixture models for the three vertical lines indicated in Figure 5.2a and 5.2b are shown in Figures 5.2d-f. In each plot, the observed distribution in black together with the mixture components (green and red, respectively, densities scaled to their estimated fractions) is shown. For quality control, the sum of both scaled component distributions is shown in blue as well as the remaining distribution after subtracting the fitted background from the observations as dashed red line. Note that the observed distribution itself is normal for the low intensity window in Figure 5.2d, but starting from intensity values of ≈ 5 , the distribution is indeed a mixture of two normal distributions.

We performed this running window approach for all previously published RIP-chip experiments from Dölken et al. [2010] as well as for two additional replicates of the control cell line DG75-eGFP and the EBV infected cell line Jijoye, respectively. For the additional DG75-eGFP

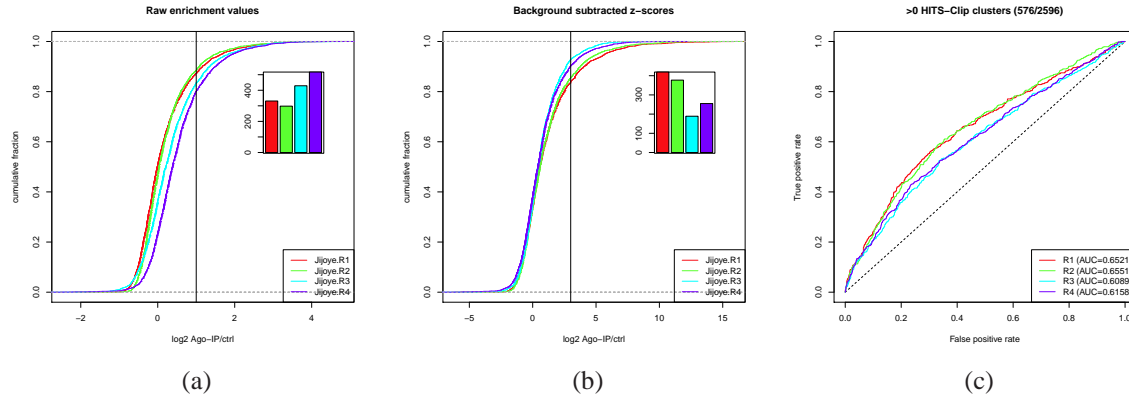


Figure 5.3: Background subtraction is necessary. Figures 5.3a and 5.3b show the enrichment distributions of expressed genes in the four Jijoye RIP-chip replicates. Raw enrichment values indicate that the IPs of replicates three and four were more efficient based on the number of genes enriched more than two fold (see corresponding inset). After background subtraction however, replicates one and two show a larger fraction of enriched genes. Replicates one and two have significantly better correspondence to the HITS-CLIP experiment performed in Jijoye, which shows the need for background subtraction.

replicates, the microarrays showed poor RNA quality and applying our running window approach to these bad quality experiments indeed did not yield a mixture model (see Figure 5.2c). Thus we can apply our method also for filtering poor quality experiments from a dataset and we excluded the two additional DG75-eGFP replicates from further analyses accordingly.

We also noticed that the background distribution is not the same over the whole spectrum of intensity values. Therefore, we computed z-scores for each gene using mean and standard deviations obtained from the running window approach. This is very similar to well known nonlinear normalization techniques [Yang et al., 2002], with the difference that the model for normalization is not fitted to all data but only to the background.

5.4.2 Determining microRNA targets

Computing z-scores from the raw enrichment values based on the fitted background distribution can be interpreted as a subtraction of this background. Note that the background here does not consist of the unexpressed genes, but of the expressed but not targeted genes. This background subtraction step can have a great effect: For the four Jijoye RIP-chip replicates, we observe that without subtraction, it seems that the IPs of replicates three and four were more efficient than of the other two replicates, since there are more genes enriched more than two fold, for instance (see Figures 5.3a and 5.3b). However, after background subtraction replicates one and two show a larger fraction of enriched genes.

Obviously, if an IP was more efficient than another, its induced ranking of genes will better predict a gold standard of microRNA targets. HITS-CLIP is an experimental technique that

is able to identify target sites of microRNAs with high confidence [Chi et al., 2009]. Thus, using the publicly available HITS-CLIP data for Jijoye [Riley et al., 2012a] we can construct a gold standard by taking all genes as true targets that have at least n HITS-CLIP target sites. Independent on the choice of n , replicates one and two induce rankings that are in better agreement with HITS-CLIP data (see Figure 5.3c for $n = 1$), and thus, background subtraction is a necessary step. We also repeated this analysis using in-house, unpublished PAR-CLIP data for Jijoye ($\sim 14,000$ sites on ~ 5500 genes, will be published elsewhere) leading to the same conclusions (data not shown).

Furthermore, we propose that the fitted background distribution allows to compute valid false discovery rates (FDRs) for microRNA targets. For a cutoff c , the FDR is defined as the expected fraction of nontarget genes. Obviously, a nontarget gene should contain less HITS-CLIP target sites than target genes on average. If we compute the average number of HITS-CLIP target sites per gene for the set of targets defined by cutoff c on the RIP-chip data, the dependence on the corresponding FDR should thus be linear with a negative slope: For instance, if the FDR is twice as high, we expect twice as many nontarget genes. Therefore, the average number of HITS-CLIP target sites per gene should decrease by a factor that is dependent on the true average number of HITS-CLIP targets sites per target gene and nontarget gene. For all four replicates the plot of the FDR against the fraction of HITS-CLIP target sites per gene is roughly a straight line (see Figure 5.4) and even if the enrichment/z-score distributions are quite different, the slopes and intercepts of linear fits to all four plots are very similar to each other (see Figure 5.5).

This also allows us to estimate the average number of HITS-CLIP target sites per target gene and nontarget gene by taking the value of the linear fit at FDR=0% and FDR=100%, respectively. Based on the RIP-chip data as a reference, we can estimate that HITS-CLIP produces ≈ 0.8 target sites per expressed target gene and ≈ 0.2 target sites per nontarget gene (see Figure 5.5).

5.4.3 Taking replicates into account

As indicated above, IP efficiencies between replicate experiments may be very different from each other. These differences introduce bias into such a dataset and no RIP-chip study known to us has properly accounted for that. Note that our mixture model approach also cannot remove this bias from RIP-chip data. The problem becomes obvious when we visually inspect scatterplots across replicate enrichment values/z-scores.

For replicates one and two of our Jijoye RIP-chip data, the main cloud of target genes roughly scatters around the main diagonal in Figure 5.6a, whereas for the comparison of replicates one and three, the diagonal is quite far away from the main cloud (Figure 5.6b). The canonical way for summarizing replicates is to take the unweighted mean of the enrichment values/z-scores. This can geometrically be interpreted as an orthogonal projection onto the diagonal vector $d = (0.25, 0.25, 0.25, 0.25)$ and measuring the distance of the projected point to the origin. Thus, all four-dimensional points lying on any hyperplane orthogonal to d would get the same summary value. Such a hyperplane would not cut the main cloud of target genes in the scatter plot of replicates one and three orthogonally, which is only a consequence of different IP efficiencies.

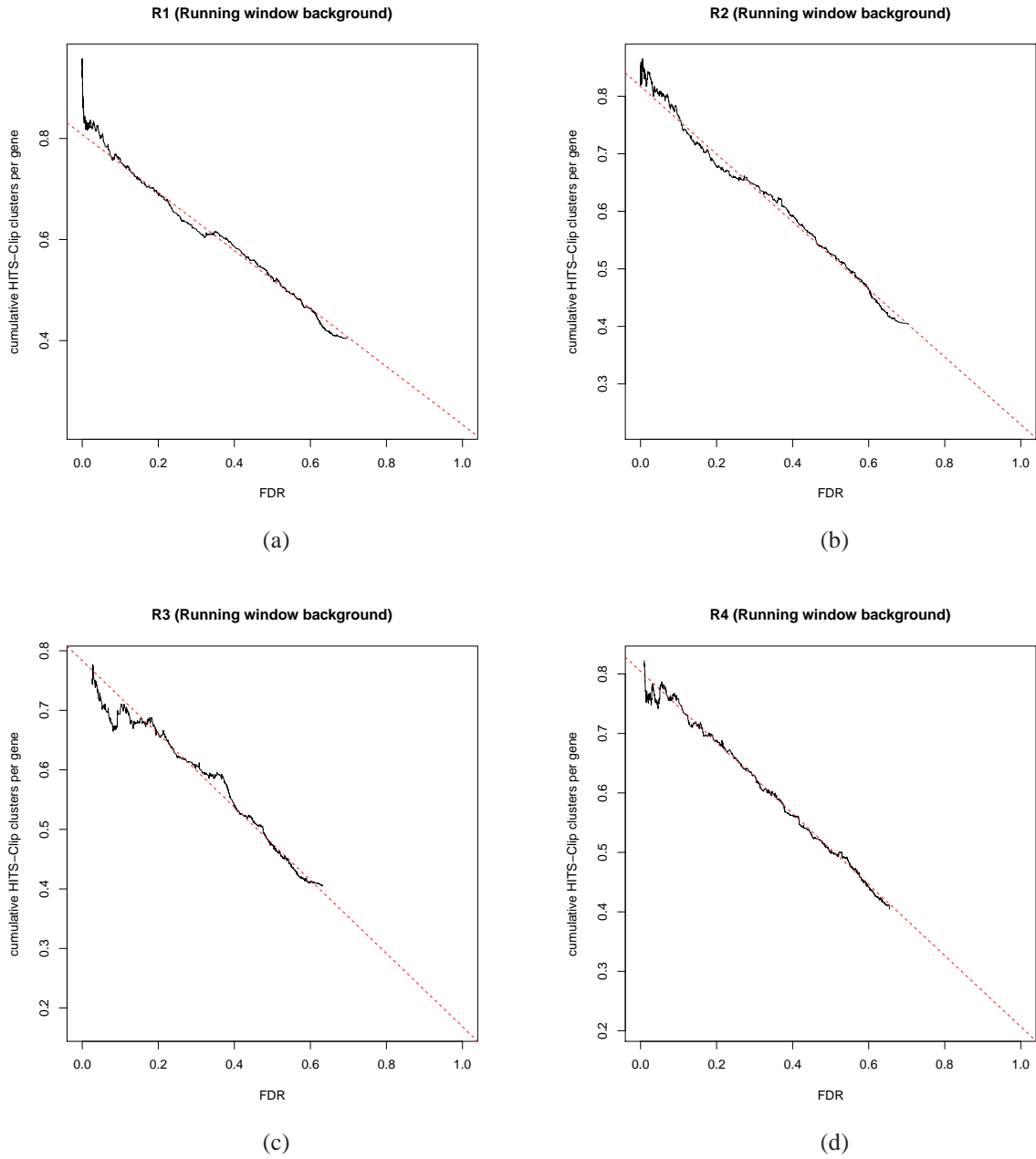


Figure 5.4: Computed FDRs are valid. The false discovery rates of the four Jijoye replicates are plotted against the average number of HITS-Clip clusters per gene. All four show a roughly linear behavior suggesting that the FDR is valid. Furthermore, linear fits to each of the plots are very similar to each other, despite of quite different z-score distributions (see Figure 5.3b), which allows us to estimate the average number of HITS-Clip target sites per RIP-chip target and background gene (see main text for further details).

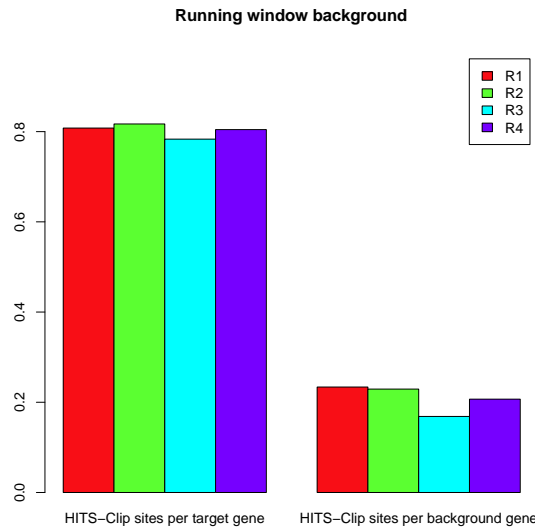


Figure 5.5: Average number of HITS-Clip target sites per RIP-chip background and target gene. The linear fits at FDR=0% and FDR=100% in Figure 5.5 are estimates for the number of HITS-Clip target sites in RIP-chip target genes and background genes, respectively. Even if the IP efficiencies were quite different (see Figure 5.1), the estimates of ≈ 2.8 target sites per expressed target gene and ≈ 1.7 per expressed background gene are remarkably similar.

However, the first principal component of this point cloud defines such orthogonal hyperplanes and we use the components of the corresponding rotation vector to compute a weighted mean accounting for all linear effects of differing IP efficiencies.

We can evaluate this additional step again by using the HITS-CLIP data as reference. We consider the differences between each normalized summary value and the corresponding unnormalized value. The difference for HITS-CLIP sites containing genes is statistically significantly greater than for other genes ($p < 10^{-14}$, Kolmogorov-Smirnov test, see Figure 5.6c), and the more HITS-CLIP targets sites are found for a gene, the more pronounced is its positive change.

5.4.4 Determining differential microRNA targets

In order to find differential microRNA targets, experiments of different conditions must be compared, i.e. genes must be identified, that are more enriched in one condition in comparison to the other. Obviously, a similar problem as in the summarization of replicates plays a role: How can we account for differing IP efficiencies if we compare four replicates of the EBV infected cell line Jijoye to the two replicates of the control cell line DG75-eGFP?

We can extend our method for summarizing replicates to the differential problem: The first principal component corresponds to the direction of greatest variance, which is the direction of common targets under the assumption that there are enough common targets (both are B cell lines). Differential microRNA targets exclusive to Jijoye should have positive enrichment

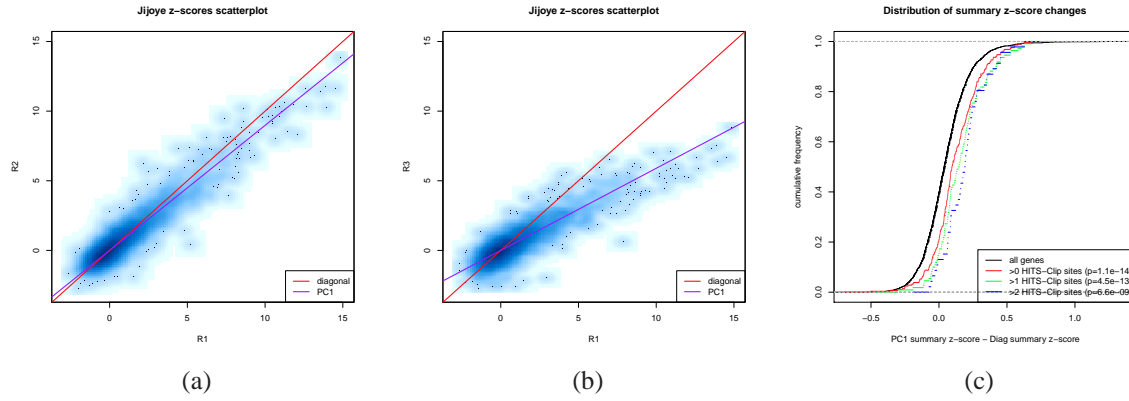


Figure 5.6: Differing IP efficiencies require normalization before computing summary values. In contrast to replicates one and two of the Jijoye RIP-chip experiments, where IP efficiencies are very similar (see Figure 5.6a), replicate three is different (see Figure 5.6b). Normalizing replicates using the first principal component (PC1) significantly improves the summary z-score with respect to HITS-CLIP data as reference: The difference of the normalized score to the unnormalized score is significantly greater for genes with HITS-CLIP target sites (colored distributions) in comparison to all differences (black distribution). The improvement is even more pronounced for genes with multiple HITS-CLIP target sites (see Figure 5.6c).

values in the Jijoye RIP-chip replicates and smaller values in DG75-eGFP. These targets induce variance into the corresponding direction of the six-dimensional space, such that the second principal component corresponds to the IP efficiency normalized direction of differential targets (see Figure 5.7a).

In order to compare the PC2 normalized differential enrichment values to the unnormalized differential enrichment (i.e. subtract the enrichment mean of DG75-eGFP from the mean of Jijoye), we exploit the fact that microRNAs are able to downregulate expression of target mRNAs [Bartel, 2009; Guo et al., 2010] and that mRNA levels were measured as well in the RIP-chip experiment: x fold downregulated genes get consistently and significantly higher scores after normalization as compared to all other genes, independent on the choice of x (see Figure 5.7). Thus, after normalization, significantly more RIP-chip targets are downregulated than without normalization (independent on the particular threshold used to define RIP-chip targets and downregulated genes).

5.5 Discussion

A similar approach to our gaussian mixture modelling (GMM) has already been used in [Mukherjee et al., 2009], however, GMM was applied to summarized enrichment values and log odds ratios (LOD scores) were computed as the ratio of the two scaled mixture components. LOD scores were then used in two different ways: First, they were directly subjected to gene

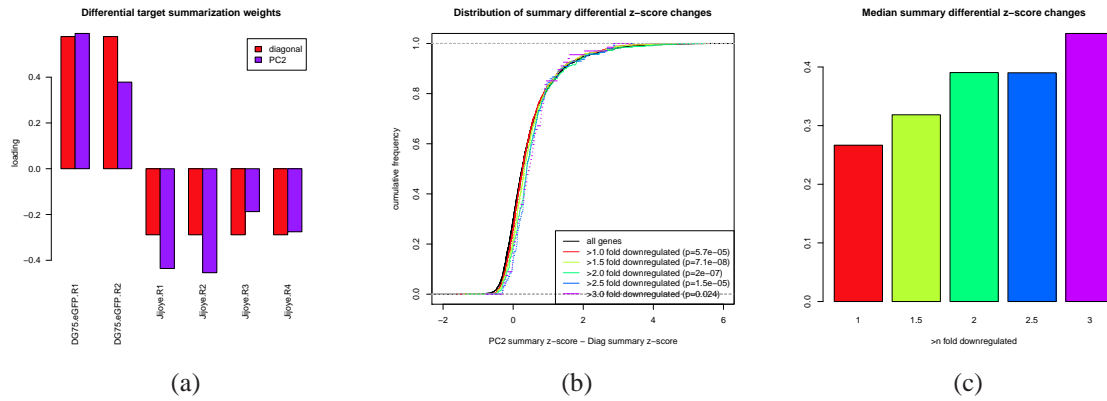


Figure 5.7: Differing IP efficiencies require normalization before computing differential targets. The second principal component in Figure 5.7a is able to discover the experimental structure. Its loadings can be used as weights to compute a differential enrichment value that is normalized for different IP efficiencies, in contrast to the standard way of subtracting the mean log enrichment in DG75-eGFP from the mean log enrichment in Jijoye (corresponding to weights indicated in red). The difference distribution of the normalized differential enrichment values and the unnormalized ones is shown in Figure 5.7b. Differential targets are expected to be downregulated, and indeed, the difference is significantly greater than background for downregulated genes. As illustrated in Figure 5.7c, this effect is more pronounced the higher the downregulation is.

set enrichment analysis (GSEA) [Subramanian et al., 2005], where they have no advantage over directly using enrichment values (since the weighted Kolmogorov-Smirnov statistic used for GSEA is non-parametric and only sensitive to the ranking of the genes). Second, the authors used a cutoff of $\text{LOD} > 0$ in order to define a set of targets. However, choosing a cutoff based on the LOD is still arbitrary and not statistically meaningful in contrast to our false discovery rates. If we used the LOD to define a cutoff, we would get FDRs ranging from 5% to 15% in our experimental dataset.

We could show that our refined mixture modelling approach has several advantages: First, it allows us to filter unexpressed genes. When comparing two conditions (e.g. virus infected cells expressing viral microRNAs vs. non-infected cells) expression of a gene targeted by cellular microRNAs below the detection limit of the microarray in one but not the other cell line would result in the misinterpretation of this to be a target of the viral microRNAs. Second, for experiments with poor IP efficiency, we observed extremely poor distance and asymmetry scores over the whole intensity range and could remove these bad replicates from further analyses. Third, it helps to compare experiments to each other (see Figure 5.3c) and finally, we can compute valid FDRs.

Furthermore, the comparison of the RIP-chip FDRs to HITS-CLIP data revealed important properties of both the RIP-chip and HITS-CLIP techniques: Both are designed to identify microRNA targets and naturally, they agree significantly ($p < 2.2 \times 10^{-16}$, Kolmogorov-Smirnov test), a fact that is also reflected in the negative slope of the linear fit to the FDR against sites

per gene plot. However, the agreement is not perfect and we estimate an average number of ≈ 0.8 HITS-CLIP target sites per RIP-chip target gene and of ≈ 0.2 per RIP-chip nontarget gene. Even if HITS-CLIP target sites may be erroneous and the CLIP techniques may implicate additional bias [König et al., 2010; Kishore et al., 2011], we expect that not all of the ≈ 0.2 sites per background gene are true errors: Such inconsistencies may be due to differing experimental steps (e.g. different antibodies used for IP) or due to differences the Jijoye cell cultures have accumulated in the two laboratories since the cell line has been established. Also, since HITS-CLIP does not control for target mRNA abundance, it may find several weak sites on highly expressed genes that are biologically irrelevant (i.e. not contributing significantly to regulation of its expression). Such a gene should not be enriched in a RIP-chip experiment and could explain many cases of HITS-CLIP sites on background genes. Thus, even if CLIP techniques have several advantages (e.g. they are able to identify target sites instead of target genes), RIP-chip is still a useful complementary method.

The second, novel method introduced in this paper is to use principal components to normalize for different IP efficiencies. Evaluation using HITS-CLIP data or the differential expression of target genes shows that the normalization improves results significantly. The normalization proposed can only account for linear bias between experiments. This very lenient normalization appears to be sufficient, since affine offsets are already removed by the mixture model approach and nonlinear effects are not recognizable in a visual inspection.

Our proposed methods do not include a way to compute statistical significance, e.g. like a t-test for standard differential gene expression experiment. However, this can be accomplished in a straight-forward way, since all available tests could directly be used after our linear normalization has been applied to a dataset.

5.6 Conclusion

In this paper we presented methods we developed to analyze RIP-chip data. In comparison to standard differential gene expression experiments, the additional immunoprecipitation step introduces special requirements for the data analysis. First, we use gaussian mixture modelling (GMM) to determine biologically significant target genes, and second, we use a linear normalization technique based on principal component analysis to remove bias introduced by the immunoprecipitation. The evaluation of both methods using independent data showed a significant improvement in comparison to standard approaches: The background of not enriched genes can be removed, valid FDRs can be calculated and the comparability of both replicates and differential experiments is improved.

Chapter 6

Widespread context-dependency of microRNA-mediated regulation

Motivation: *In the previous two chapters, I introduced two methods for raw data analysis of PAR-CLIP and RIP-Chip data, respectively. As indicated in the introduction (see section 1.1.1), raw data analysis is an important step in systems biology and converts raw data from an experiment to biological information. However, in order to understand a biological system as a whole, this information must be interpreted. In this chapter, I describe how microRNA related data can be interpreted with respect to different cellular contexts. There were two specific reasons that lead me to consider this specific aspect of microRNA-mediated regulation: First, one of the first questions that came up when the PAR-CLIP data from our collaboration partners became available was how big the overlap between our dataset and another already published KSHV related PAR-CLIP dataset is [Gottwein et al., 2011]. This question originally was about data quality and sounds quite easy to answer at first. However, in fact it is not, which has to do with the way how an overlap should be defined for PAR-CLIP datasets and what the implications are of any size of overlap. Second, during that time, the main phase of the ENCODE project was published in several papers in Nature, Genome Research and Genome Biology and one of the most intriguing results was that transcriptional regulation is heavily dependent on the cellular context. So the main question was, whether and to which extent this is also true for microRNA-mediated regulation. Importantly, in our project, various datasets have been generated that allowed me to investigate and resolve this question.*

Publication: *This chapter has been submitted for publication [Erhard et al., 2013c]. Here, I adapted the layout and restructured parts of the text to incorporate important parts of the Supplementary material of the submitted manuscript into this chapter.*

My contribution: *I analyzed PAR-CLIP, RIP-Chip and mass spectrometry data and came up with the idea of context-dependency of PAR-CLIP target sites. I carried out all computational and statistical analyses, produced plots and wrote the paper.*

Contribution of co-authors: *Lukasz Jaskiewicz and Mihaela Zavolan performed PAR-CLIP experiments, Georg Malterer, Diana Lieber and Jürgen Haas provided SILAC measurements,*

*Lars Dölken provided RIP-Chip and 4sU-tagging datasets and helped to revise the manuscript.
Ralf Zimmer supervised the work and helped to revise the manuscript.*

6.1 Abstract

Gene expression is regulated in a context-dependent, cell-type specific manner. Condition-specific transcription is dependent on the presence of transcription factors (TFs) that can activate or inhibit its target genes (global context). Additional factors such as chromatin structure, histone or DNA modifications also influence the activity of individual target genes (individual context). The role of the global and individual context for post-transcriptional regulation has not systematically been investigated on a large-scale and is poorly understood. Here we show that global and individual context-dependency is a pervasive feature of microRNA-mediated regulation. Our comprehensive and highly consistent dataset from several high-throughput technologies (PAR-CLIP, RIP-Chip, 4sU-tagging and SILAC) provides strong evidence that context-dependent microRNA target sites (CDTS) are as frequent and functionally relevant as constitutive target sites (CTS). Furthermore, we found the global context to be insufficient to explain the CDTS and that RNA binding proteins provide individual context that is an equally important factor. Our results demonstrate that similar to TF-mediated regulation, global and individual context-dependency are prevalent in microRNA-mediated gene regulation implying a much more complex post-transcriptional regulatory network than currently known. The necessary tools to unravel post-transcriptional regulations and mechanisms need to be much more involved and much more data will be needed for particular cell types and cellular conditions to understand microRNA-mediated regulation and the context-dependent post-transcriptional regulatory network.

6.2 Introduction

Regulation of gene expression is highly context-specific. The ENCODE project [Consortium, 2012b] provided convincing evidence that whether or not a specific transcription factor (TF) binds to a specific binding site (TFBS) is not only dependent on the sequence of the binding site but also on its chromatin state [Wang et al., 2012b], on DNA methylation [Wang et al., 2012a], on other DNA binding factors [Yanez-Cuna et al., 2012] and numerous additional factors, which are difficult to measure and predict. All these factors form the so-called *cellular context* that influences the expression level of genes.

Gene expression is not only regulated at the level of transcription but also post-transcriptionally in various ways of which regulation mediated by microRNAs is one of the most prevalent [He and Hannon, 2004]. MicroRNAs are 20-24 nt long non-coding RNAs that have been found in animals and plants. They play a pivotal role in development, tumorigenesis, the immune system and during viral infections (for a review see Bartel [2004]). Within the RNA induced silencing complex (RISC), microRNAs are responsible for target recognition by binding to target sites, often located in the 3'-UTR of mRNAs. This is predominantly mediated by the so-called seed region (nucleotides 2-8 of the microRNA). In general, RISC causes downregulation of the target mRNA either by inhibiting translation or promoting degradation [Bartel, 2009]. Neither the exact mode of binding nor the mechanisms of downregulation are completely understood [Djuranovic

et al., 2011; Eulalio et al., 2008; Guo et al., 2010; Kozak, 2008; Mishima et al., 2012; Meijer et al., 2013].

Computational prediction of microRNA targets is a difficult task [Thomas et al., 2010; Sethupathy et al., 2006; Ritchie et al., 2009]. This is a consequence of the low specificity of seed matches alone: There are several lines of evidence suggesting that additional factors such as target site location [Grimson et al., 2007], additional basepairing at the microRNA 3' end [Brennecke et al., 2005], target site accessibility [Kertesz et al., 2007], other RNA binding proteins [Jacobsen et al., 2010], microRNA and mRNA copy numbers [Ben-Moshe et al., 2012], additional unknown factors or interplay between any of these play important roles in distinguishing functional from non-functional target sites. Interestingly, several of these additional factors are not static but may change dynamically: For instance, dependent on which RNA binding proteins are expressed at what level in a given cell-type, RISC may or may not bind at a certain binding site. Similarly to transcription factors, microRNAs are therefore likely to exhibit their regulatory function in a context-dependent manner.

Several examples of context-specific microRNA-mediated regulation can be found in the literature (for a review see Pasquinelli [2012]). Bhattacharyya et al. [2006] identified the RNA binding protein HuR as a derepressor for miR-122 regulation of the CAT-1 mRNA. In normal hepatocarcinoma cells CAT-1 is repressed by a miR-122 target site in its 3'-UTR. Under different stress conditions, HuR is released from the nucleus into the cytoplasm which abolishes CAT-1 repression. The exact mechanism however remains unclear. Intriguingly, HuR has also been implicated in activating a target site of the microRNA let-7 in the 3'-UTR of MYC [Kim et al., 2009a] which indicates that HuR can both induce and prevent microRNA-mediated regulation. In addition to HuR, DND1 [Kedde et al., 2007] and Pumilio-1 [Kedde et al., 2010] have also been identified to influence microRNA regulation. There may be other RNA binding proteins that interfere with or facilitate microRNA/target interactions.

These examples illustrate that the presence of a functional target site is not sufficient for regulation. It may be active under certain conditions but non-functional in a different context. Presently, our knowledge about context-dependent microRNA-mediated regulation is based on few examples and the underlying molecular mechanisms are poorly understood.

By immunoprecipitation of microRNA/target complexes using monoclonal antibodies to RISC components followed by high-throughput sequencing of the protein-protected microRNA target sites, the complete targetome of cellular and viral microRNAs has become accessible. More than 10,000 putative microRNA binding sites, so called clusters, are obtained in a single HITS-CLIP (high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation) or PAR-CLIP (photoactivated ribonucleotide-enhanced crosslinking immunoprecipitation) experiment. Although the annotation of the responsible microRNA to an identified cluster still leaves room for improvement, more than 75% of microRNA target interactions can be correctly annotated thereby allowing in-depth analyses of microRNA regulatory networks [Gottwein et al., 2011; Skalsky et al., 2012; Haecker et al., 2012; Riley et al., 2012b].

To study context specific microRNA-mediated regulation, we generated Ago2-PAR-CLIP data from two human B-cell lines. In addition, we re-analyzed two recently published sets of Ago2-PAR-CLIP data from two different human B-cell lines [Gottwein et al., 2011]. These four cell lines represent different stages of B-cell development and are either infected by Kaposi's

sarcoma-associated herpesvirus (KSHV), co-infected by both KSHV and Epstein-Barr-Virus (EBV) or not infected. Thus, each cell line provides a distinct context for microRNA-mediated regulation. All datasets were re-analyzed using a new algorithm called PARma [Erhard et al., 2013a]. PARma considers the topology of the microRNA/target interaction and the position of UV-light induced cross-links in more detail than state-of-the-art methods and provides quality control scores for both, the identification of microRNA target site clusters and for the annotation of the interacting microRNA to these sites. For two of these four cell lines, we generated three additional data sets including RIP-Chip, 4sU-tagging-derived RNA half-lives and large-scale SILAC-based proteomics. This allowed us to comprehensively analyze the effect of context-dependent microRNA/target interactions on the recruitment of the target mRNAs to Argonaute-2 complexes, on target RNA stability and on target protein levels. By considering viral as well as host microRNAs, we investigated both microRNA/target interactions that coevolved within a species as well as interactions of an exogenous microRNA with endogenous target sites. The results provide compelling evidence that context-dependency of microRNA-mediated regulation is not restricted to a few examples but is a widespread and general feature of post-transcriptional regulation mediated by both cellular and viral microRNAs.

6.3 Results

6.3.1 Differential analysis of PAR-CLIP data

To comprehensively study regulation of cellular gene expression by both cellular and Kaposi's sarcoma-associated herpesvirus (KSHV)-encoded microRNAs, we applied Ago2-PAR-CLIP to two human B-cell lines, the body cavity based lymphoma cell line BCBL1, which is latently infected with KSHV, and the Burkitt lymphoma cell line DG75, which is KSHV negative. Applying PARma [Erhard et al., 2013a] with stringent criteria (see the Methods), we identified 15,577 clusters, 12,333 of which mapped to known transcripts (Ensembl v60).

In order to assess the quality of the PAR-CLIP datasets, we first computed the positional distribution of all target sites in mRNAs (Figure 6.1a). Target sites of viral microRNAs shared the well described features of cellular microRNA target sites: They preferentially bind to the 3' untranslated region (3'-UTR) and rarely to the 5'-UTR of transcripts [Grimson et al., 2007; Hafner et al., 2010]. Within the 3'-UTR, target sites tend to accumulate at the very beginning, i.e. immediately after the stop codon, and at the transcript end, i.e. immediately upstream of the poly-A tail [Grimson et al., 2007]. We furthermore checked the accuracy of the microRNA assignment to target sites by confirming that virtually no reads mapped to KSHV microRNA target sites in the KSHV negative cell line DG75 (a feature that is not used by PARma to assign microRNAs; see Figure 6.1b). The few instances with random reads in DG75 may nevertheless be bona-fide KSHV microRNA target sites: As we observed random reads spread across a multitude of transcripts at low frequency, these reads presumably result from infrequent unspecific immunoprecipitates or insufficient removal of background total RNA rather than microRNA-specific signatures. This is further supported by a significantly lower frequency of T to C conversions and lower consistency across replicates for these reads (Figure 6.1b).

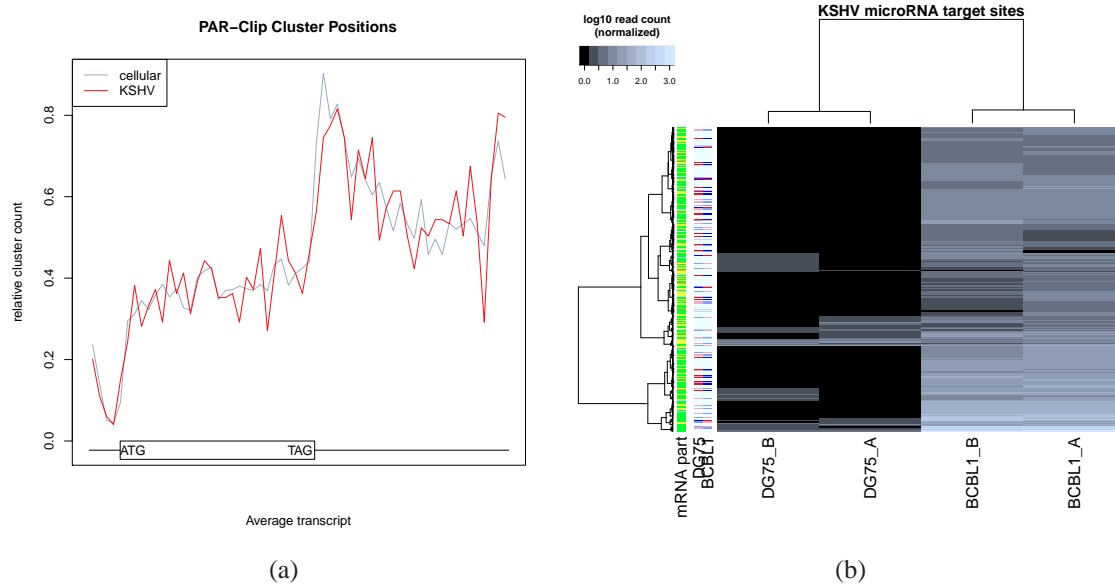


Figure 6.1: Validation of PAR-CLIP experiments. In Figure 6.1a, the distribution of relative positions of target sites on mRNAs is shown. The x-axis represents the average length of 5' untranslated regions (5'UTR), of the coding regions (CDS) and of the 3' untranslated regions of all transcripts with at least one PAR-CLIP cluster. Each transcript was divided into 60 bins and the relative frequency of target sites falling into each bin is shown on the y-axis. The data clearly illustrate the preferences of target site in the 3'UTR as compared to CDS and 5'UTR. Viral microRNAs have the same preferences as cellular microRNAs. In Figure 6.1b the normalized number of reads in each cluster (rows) for each of the independent PAR-CLIP experiments (columns) is shown for KSHV microRNA target sites in the four PAR-CLIP libraries. KSHV negative cell lines (columns 1 and 2) almost exclusively have no reads, whereas for KSHV positive cell lines, dozens to hundreds of reads are observed per target site. Replicates are highly correlated indicating high reproducibility. The additional annotations on the left side indicate the part of the transcript, where a cluster is located (orange: 5'-UTR; yellow: coding; green: 3'-UTR; gray: not located on known mRNA) and the expression of the transcript in all experiments (red, at least 2-fold lower expression than the mean expression value for this transcript across all experiments; light red, at least 1.4-fold lower expression than the mean; light blue, at least 1.4-fold higher expression; blue, at least 2-fold higher expression).

We further validated our PAR-CLIP dataset using published data for the same cell lines: (i) PAR-CLIP targets are highly consistent with RIP-Chip data (Dölken et al. [2010], Figures 6.2a and b), (ii) KSHV microRNA targets are selectively enriched in BCBL1 and not DG75 in the RIP-Chip experiments (compare Figures 6.2a and b) and (iii) PAR-CLIP target sites lead to a measurable reduction of target mRNA half-lives (Figure 6.2c).

To be able to perform a more in-depth analysis on KSHV microRNA targets in human B-cells, we also included recently published PAR-CLIP data from two additional B-cell lines, namely BC1

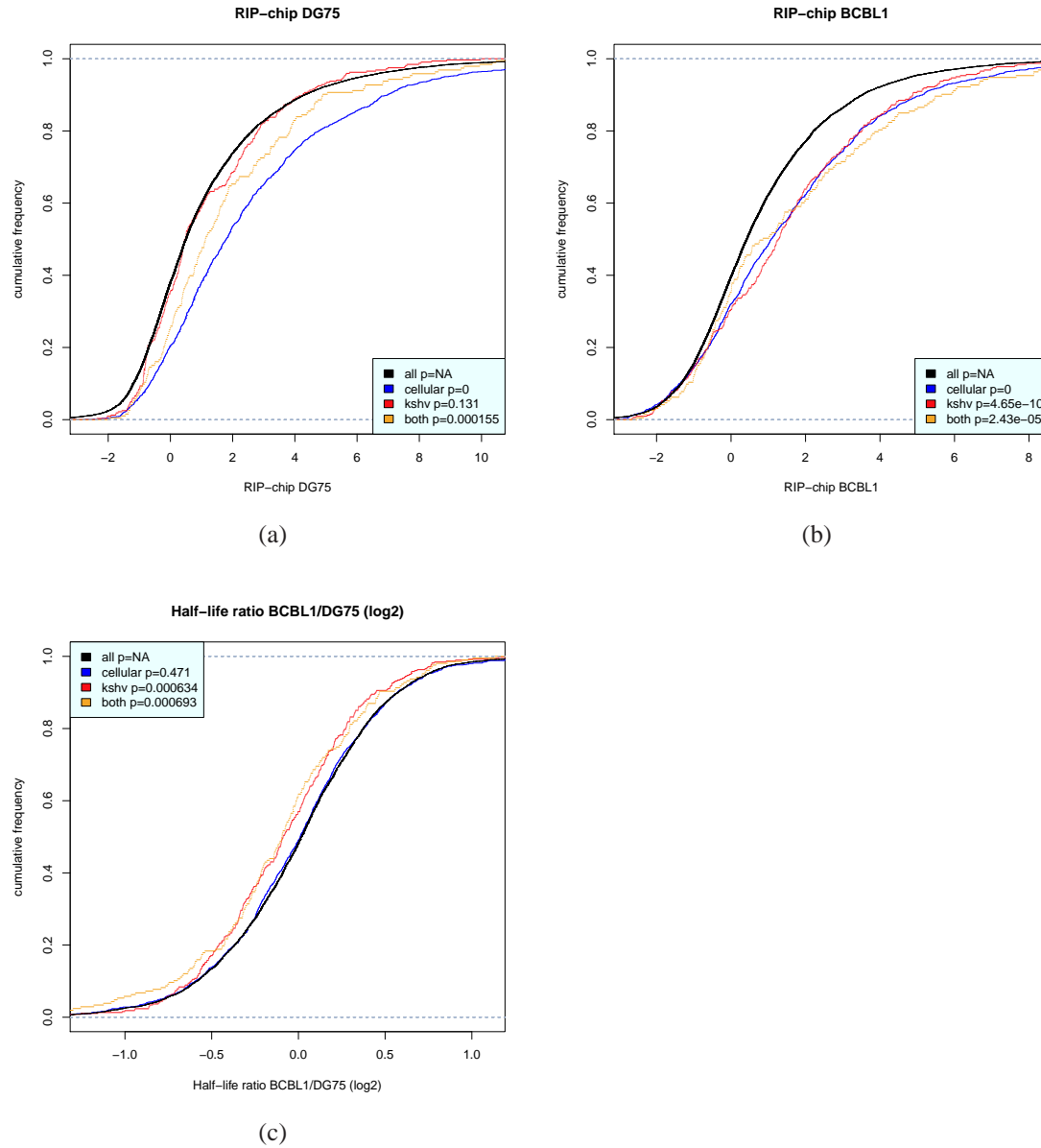


Figure 6.2: Comparison of PAR-CLIP experiments with available datasets. Figures 6.2a and 6.2b show the log₂ RIP-Chip enrichment distributions of mRNAs only containing target sites of cellular microRNAs, only containing KSHV microRNA target sites and containing target sites from both cellular and KSHV microRNAs in the uninfected cell line DG75 and the KSHV positive cell line BCBL1, respectively. KSHV targets are enriched in BCBL1 but not in DG75. In Figure 6.2c, the mRNA half-life ratios are shown for the same sets of genes as in Figure 6.2a and b. The half-life of mRNAs with KSHV target sites is significantly reduced in BCBL1.

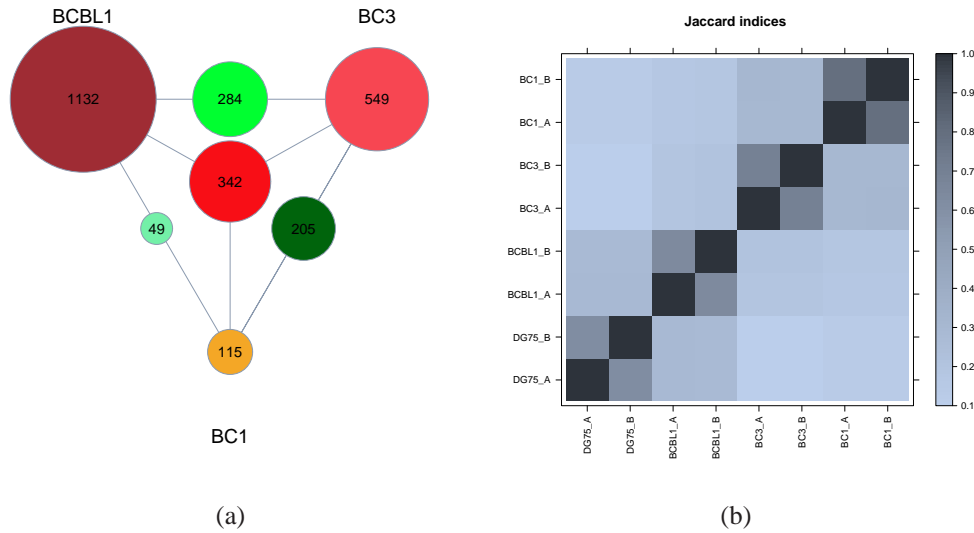


Figure 6.3: Comparison of PAR-CLIP datasets. Figure 6.3a illustrates the number of target sites observed only in individual cell lines (outermost labeled circles), in two cell lines (circles on the edges between cell lines) and in all three cell lines (center circle), for KSHV microRNA target sites. Relatively few target sites appear to be active in multiple cell lines. Figure 6.3b summarizes all pairwise overlaps for all clusters in all datasets. The Jaccard index (J) is the number of clusters in the intersection divided by the total number of clusters in any of the two experiments. Jaccard indices of about 70% for all replicate measurements indicate high reproducibility, whereas comparisons across cell lines show relatively low overlap ($J < 40\%$) (see also Figure 6.4).

and BC3 [Gottwein et al., 2011]. We re-analyzed all datasets using PARma, which yielded 21,628 clusters, 16,425 of which mapped to known transcripts. Intriguingly, the overlaps of targets sites of both ubiquitously expressed cellular and KSHV microRNAs were surprisingly small (Figure 6.3b and 6.3a). Such extreme differences of called target sites may be due to experimental bias or context-dependency, i.e. a major fraction of microRNA target sites is only active in some of the cell lines considered.

6.3.2 Technical bias

When analyzing high-throughput data obtained from experiments performed in different laboratories a certain extent of differences in between given data sets can be expected. In our case, distinct clusters of target sites may also be consequences of erroneously assigned microRNAs, bias introduced by differing sample preparation methods or insufficient sequencing depth/sequencing library complexity.

Differing sample preparation methods are the most likely cause of bias and differences in microRNA targets obtained by PAR-CLIP data. As such, the RNase used to trim the Argonaute-2 protected microRNA target sequences has recently been shown to be a major source of bias

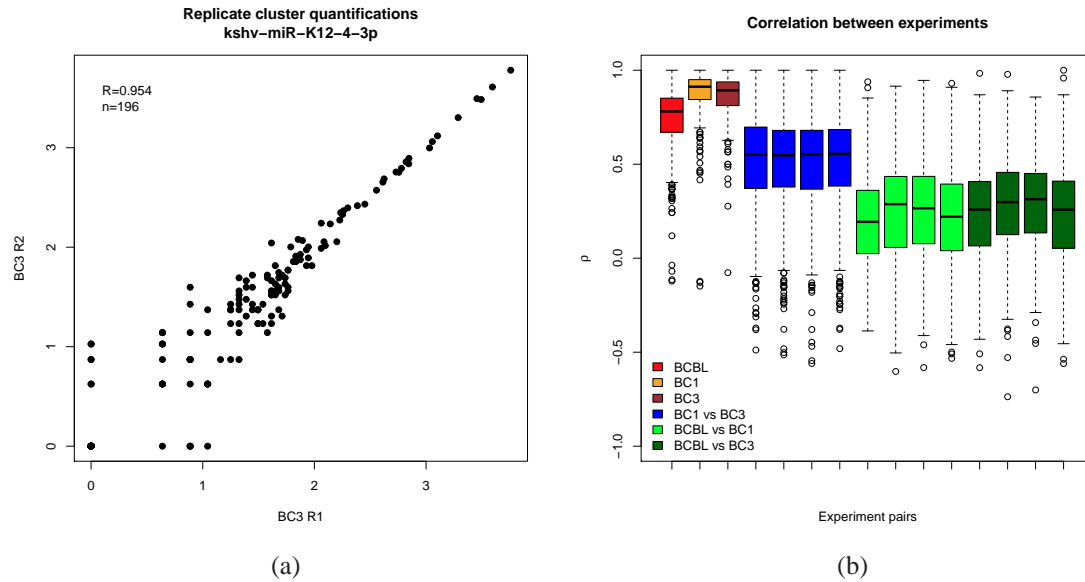


Figure 6.4: Correlations of PAR-CLIP cluster quantifications. Figure 6.4a shows the normalized quantifications of all target sites of kshv-miR-K12-4-3p as a scatter plot. Quantifications are highly correlated. In Figure 6.4b, the distributions for all microRNAs of Spearman's ρ (non-parametric correlation coefficient) are shown for each possible comparison of KSHV positive cell lines. Replicate correlations are drastically higher than between experiment correlations, indicating context specific microRNA targeting.

[Kishore et al., 2011]. Other sources may include differing immunoprecipitation efficiencies, RNase treatment times, sequencing adapters or any other slight variation in the PAR-CLIP protocol which may all result in a target site to be identified in one experiment but not in another. Such bias could be controlled for, if PAR-CLIP data for one or more cell lines were available that have been measured in multiple labs. And only considering the high correlation of replicate measurements does exclude such technical bias. Nevertheless, we can, to some extent, use the inverse argument: BC1 and BC3 were analyzed in the same lab using the same protocol. Thus, if technical bias was responsible for poor correlation and exclusive sites and not context-specific microRNA targeting, the correlation between BC1 and BC3 should be as high as for replicate experiments of either cell line. As illustrated in Figure 6.4b, this is not the case. However, the correlations between BCBL1 and BC3 are even lower than between BC1 and BC3, for instance. There may be two reasons for that: Either BC1 and BC3 are more similar to each other than BCBL1 and BC3 with respect to their cellular context for microRNA-mediated regulation or between-lab comparison of PAR-CLIP target sites is indeed influenced by technical bias to some extent. But nevertheless, technical bias cannot explain the relatively low correlations between BC1 and BC3 which provides first evidence that the observed differences may indeed not only be due to technical bias.

Insufficient saturation of PAR-CLIP libraries or sequencing depth may result in seemingly cell-type specific and thus context dependent microRNA/target interactions may also be a reason for our observations: If only a small fraction of target sites was detected (either due to poor immunoprecipitation efficiency or insufficient sequencing depth), sampling effects would play a severe role: Only due to limited sampling, a cluster may get very few or no reads in one sample and many reads in another, even if the target site is strongly associated with a microRNA in both experiments. Considering only a single experiment, this cannot be excluded by simply counting reads and computing statistical significance under a naive probabilistic model: There are many sequences that are identified multiple times, which may have two reasons: Either there was only a single RNA molecule of this sequence in the library and the amplification before sequencing gave rise to these multiply sequenced reads (indicating insufficient saturation), or there were multiple copies of such an RNA already in the library. Particularly for PAR-CLIP data, the latter is highly probable, since RNase T1, which is used in the PAR-CLIP protocol, cleaves in a sequence-specific way downstream of guanines [Pace et al., 1991]. Since the target mRNA seems highly accessible for cleavage outside of the microRNA target site, the number of possible distinct sequencing reads for a cluster is highly constrained. However, when we consider replicate measurements of targets sites for a specific microRNA, e.g. for BC3 (see Figure 6.4a), we observe that they are highly correlated (median $\rho > 0.77$ across all microRNAs for all replicate pairs, see Figure 6.4b). Therefore, all sequencing data utilized in this meta-analysis were found to be of sufficient saturation not to inflict major bias to our analysis.

6.3.3 Context-dependent target sites of KSHV microRNAs

Since technical bias cannot explain the differences of identified target sites across cell lines, we analyzed the possibility of context-dependency in microRNA-mediated regulation. Intriguingly, when we considered all target sites of a single microRNA, there was no clear correlation of target sites across cell lines (Figures 6.4 and 6.5). Instead, distinct clusters of target sites emerged, for instance several *kshv-miR-K12-4-3p* target sites that appear to be active in BCBL1 only and not in BC1 or BC3. This suggests that context-dependent microRNA-mediated regulation may be substantially more important than generally expected. Interestingly, the cellular context leading to these clusters of target sites is not solely determined by mRNA levels, which are indicated on the left side of the heatmap in Figure 6.5. Otherwise, one would expect significantly higher mRNA levels for BCBL1 specific target sites in BCBL1 than in BC1 and BC3, for instance. Additionally, there are target sites that are missing in BCBL1 and active in BC1 or BC3. Thus, not all target sites exclusively active in BCBL1 can be explained by a higher expression or activity of the respective microRNA or mRNA.

Taken together, our differential analysis of PAR-CLIP data suggests that microRNA-mediated regulation is substantially and generally dependent on the cellular context. To experimentally test this hypothesis, we employed three sets of additional high-throughput methods to investigate the consequences of context-dependent microRNA-mediated regulation. First, using RIP-Chip we tested whether context-dependent microRNA/target interactions, as found in the PAR-CLIP data, had a measurable impact on the recruitment of the target mRNA to RISC in their specific context only. And second, using microarray-based transcriptomics, including metabolic labeling

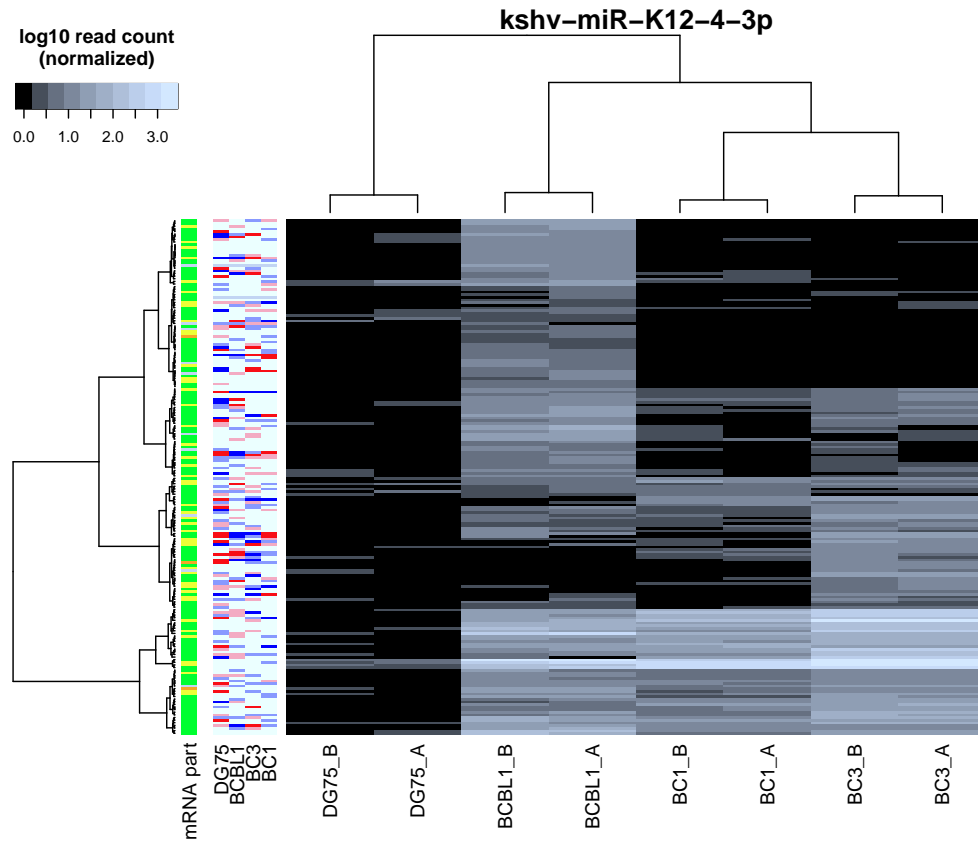


Figure 6.5: miR-K12-4-3p heatmap. The PAR-CLIP read heatmap for target sites of the KSHV microRNA miR-K12-4-3p is shown (see Figure 6.1b for more information about PAR-CLIP read heatmaps). Between KSHV positive cell lines, there is no correlation but there are distinct clusters of target sites. No obvious dependency between clusters and mRNA expression level is observable.

of RNA, and SILAC-based proteomics experiments, we tested whether such context-dependent microRNA/target interactions also have a measurable impact on mRNA half-lives and on mRNA as well as protein levels of their targets in their specific context only. All these experiments were performed by comparing DG75 to BCBL1. We selected all KSHV microRNAs that showed a KSHV specific activity pattern, i.e. where the set of target sites was depleted of reads in DG75 and included reproducible target sites of all three KSHV positive cell lines (Figure 6.6).

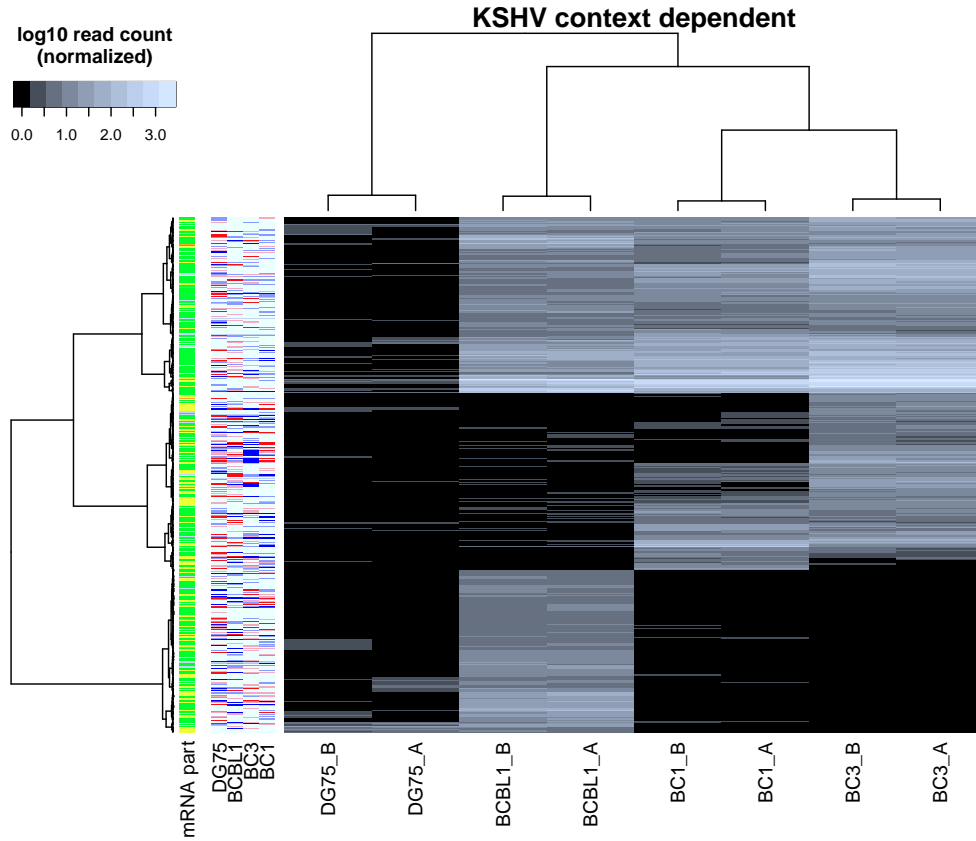


Figure 6.6: Context dependent target sites of KSHV microRNAs. The differential analysis of all sets of considered KSHV microRNAs is shown. None of the target sites has a significant amount of reads in the uninfected control cell line DG75. The top third corresponds to constitutive target sites that are active in all three KSHV positive cell line ($n = 162$), the middle third are target sites exclusively active in BC1 or BC3 and not in BCBL1 ($n = 151$) and the bottom third shows BCBL1 exclusive active target sites ($n = 151$).

Context-dependent microRNA targets are associated with RISC in a context-dependent manner

First, we looked at the recruitment of the mRNA targets of these KSHV microRNAs to Ago2-complexes. We recently employed RIP-Chip to identify KSHV and EBV microRNA targets in human B-cells [Dölken et al., 2010]. Since then, we performed two additional RIP-Chip replicates of the KSHV-positive cell line BCBL1 to perform a more solid statistical analysis [Erhard et al., 2013b].

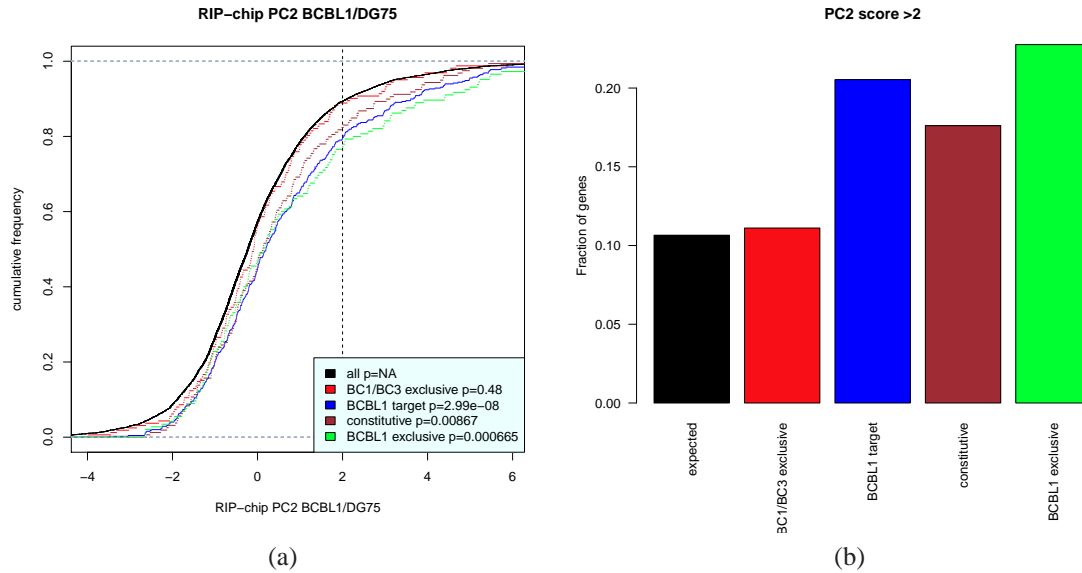


Figure 6.7: KSHV PAR-CLIP targets in RIP-Chip data. Figure 6.7a shows differential RIP-Chip enrichment scores (PC2 scores; positive values indicate higher enrichment in BCBL1 than in DG75). Generally, KSHV microRNA targets active in BCBL1 are significantly shifted towards higher values as compared to all other genes with any PAR-CLIP target site, in contrast to KSHV target sites exclusively active in BC1 or BC3 and not in BCBL1. Figure 6.7b illustrates this further: The enrichment of genes with any KSHV site, with a constitutive or a BCBL1 exclusive site over genes with BC1/BC3 exclusive sites among all genes with PC2 score > 2 is about 2-fold in all cases.

Data were normalized using principal component analysis as described [Erhard et al., 2013b] and differential enrichment values were computed for BCBL1 and DG75 as the second principal component (PC2), indicating whether an mRNA is stronger associated with RISC in BCBL1 in comparison to DG75. All PAR-CLIP target sites were mapped to genes and genes with any KSHV target site in BCBL1, with a constitutive target site in all KSHV positive cell lines and with exclusive sites in BCBL1 or BC1/BC3 were compared to all other genes with any PAR-CLIP target site as background (Figures 6.7, 6.8 and 6.9).

The differential RIP-Chip enrichment was significantly shifted towards higher values for genes with BCBL1 exclusive sites in comparison to the background ($p < 0.0007$, Kolmogorov-Smirnov test), indicating that BCBL1 exclusive target sites indeed lead to a stronger association of the target mRNA in BCBL1 to RISC. This was also true for constitutive KSHV target sites ($p < 0.009$) as well as for all KSHV target sites active in BCBL1 ($p < 3 \cdot 10^{-8}$). Moreover, BC1/BC3 specific target sites, which were not active in BCBL1, were indistinguishable from the background (Figure 6.7a). In particular, all genes with active KSHV microRNA target sites in BCBL1 showed a 2-fold enrichment of genes that are significantly (PC2 score > 2) more associated with RISC in BCBL1 than in DG75 over background genes. In contrast, genes with

KSHV microRNA target sites that are exclusively active in BC1 or BC3 and not in BCBL1 are indistinguishable from background genes (Figure 6.7b).

This provides strong evidence that a major fraction of the KSHV microRNA target sites identified by PAR-CLIP exclusively in BC1/BC3 and not in BCBL1 do not mediate a strong recruitment of their target mRNA to RISC in BCBL1, i.e. are indeed context-dependent target sites. Context-dependent microRNA/target interactions as defined by differential analysis of PAR-CLIP data can thus be confirmed using an independent RIP-Chip experiment.

Target mRNA stability is affected in a context-dependent manner

Next, we analyzed context-dependent effects of the KSHV microRNAs on target RNA stability. Since microRNAs can induce destabilization of the mRNA transcripts [Bartel, 2009], microRNA/target interactions that are active in BCBL1 should decrease the target mRNA half-life in BCBL1 as compared to DG75. Target sites inactive in BCBL1 (and only active in BC1/BC3) in contrast should not decrease mRNA half-life.

Previously, we applied metabolic labeling of newly transcribed RNA followed by microarray analysis to separate newly synthesized and pre-existing RNA [Dölken et al., 2008]. We computed RNA half-lives based on the ratios of newly synthesized to total RNA for both DG75 and BCBL1 [Dölken et al., 2010] and considered the differences in target mRNA half-lives in between BCBL1 and DG75.

Intriguingly, the mRNA half-life of KSHV microRNA targets in BCBL1 was decreased by about 20 minutes ($p < 3 \cdot 10^{-5}$) on average, whereas for KSHV microRNA targets not active in BCBL1, no significant decrease was observed (Figure 6.8a). Furthermore, the half-life difference values of BCBL1 exclusive target genes were significantly smaller than half-life difference values of BC1 or BC3 exclusive target genes ($p < 0.008$, Wilcoxon rank sum test; Figure 6.8b). Thus, context-dependent microRNA/target interactions have an impact on mRNA stability in a context-dependent manner.

Interestingly, constitutive KSHV microRNA target sites showed an even stronger decrease in the mRNA half-life than for context-dependent target sites (> 35 minutes on average, $p < 10^{-5}$). A possible explanation is that constitutive microRNA/target interactions are less susceptible to the cellular context resulting in more substantial target suppression. Therefore, constitutive interactions likely represent the most important targets for the virus.

Protein levels are differentially regulated for context-dependent microRNA targets

We now asked whether context-dependent microRNA targets are also reflected in steady-state mRNA or protein levels in two different contexts. It is important to note that protein levels in a cell depend on multiple factors, including protein half-lives and microRNA independent post-transcriptional regulation, most of which are well described to have a substantially greater impact on protein levels than generally exerted by microRNAs. Therefore, targets of the viral microRNAs may not necessarily show differential expression between DG75 and BCBL1 on protein or mRNA levels [Dölken et al., 2010]. Especially viral microRNAs are likely to

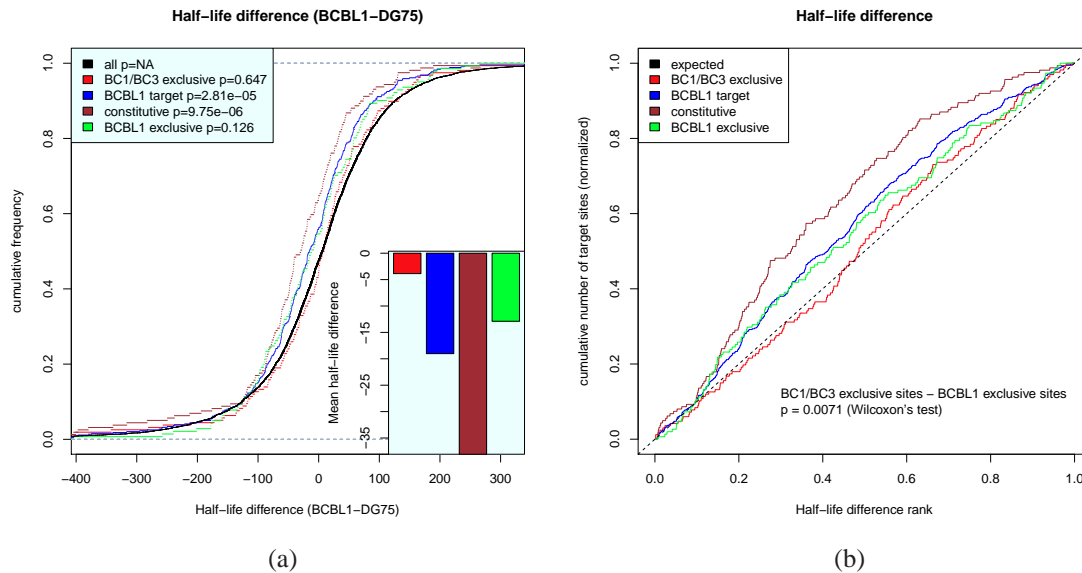


Figure 6.8: KSHV PAR-CLIP targets in mRNA half-life data. Figure 6.8a shows the distributions of half-life differences between BCBL1 and DG75 for all genes with PAR-CLIP target sites. Thus, positive values indicate a longer mRNA half-life in BCBL1 than in DG75. Genes with KSHV microRNA targets active in BCBL1 tend to have shorter half-lives in BCBL1 than in DG75. This is highly significant for all BCBL1 target genes as well as the constitutive targets but not for BCBL1 specific targets, even if their half-life is on average about 20 minutes shorter in BCBL1 than in DG75. However, KSHV microRNA targets that are inactive in BCBL1 do not show any shift in their half-lives. As illustrated in Figure 6.8b, the difference between targets active exclusively in BCBL1 is statistically significantly different from targets active exclusively in BC1 or BC3, when their ranks among all PAR-CLIP targets are considered.

counteract the cellular response to infection [Cullen, 2011; Kincaid and Sullivan, 2012] which is reflected by the fact that KSHV microRNAs target several induced genes [Dölken et al., 2010].

Indeed, when mRNA or protein levels were considered individually, no significant shift in expression fold changes was observed for any set of microRNA targets (Figures 6.9). Thus, in spite of the fact that mRNA half-lives are significantly decreased by KSHV microRNAs, there is no observable effect on steady-state levels of neither mRNAs nor proteins. However, if protein fold changes are normalized to mRNA fold changes, a small but statistically significant difference can be observed between BCBL1 specific targets and BC1/BC3 specific targets ($p < 0.01$, Wilcoxon rank sum test; Figure 6.9d). Since this normalization effectively removes all effects of mRNA levels and half-lives, this indicates that KSHV microRNAs not only have an impact on mRNA half-life in a context-dependent manner, but also on how many proteins are produced per mRNA molecule. Constitutive targets of KSHV microRNAs did not show this pattern, presumably because of their strong impact on mRNA half-lives (Figure 6.8).

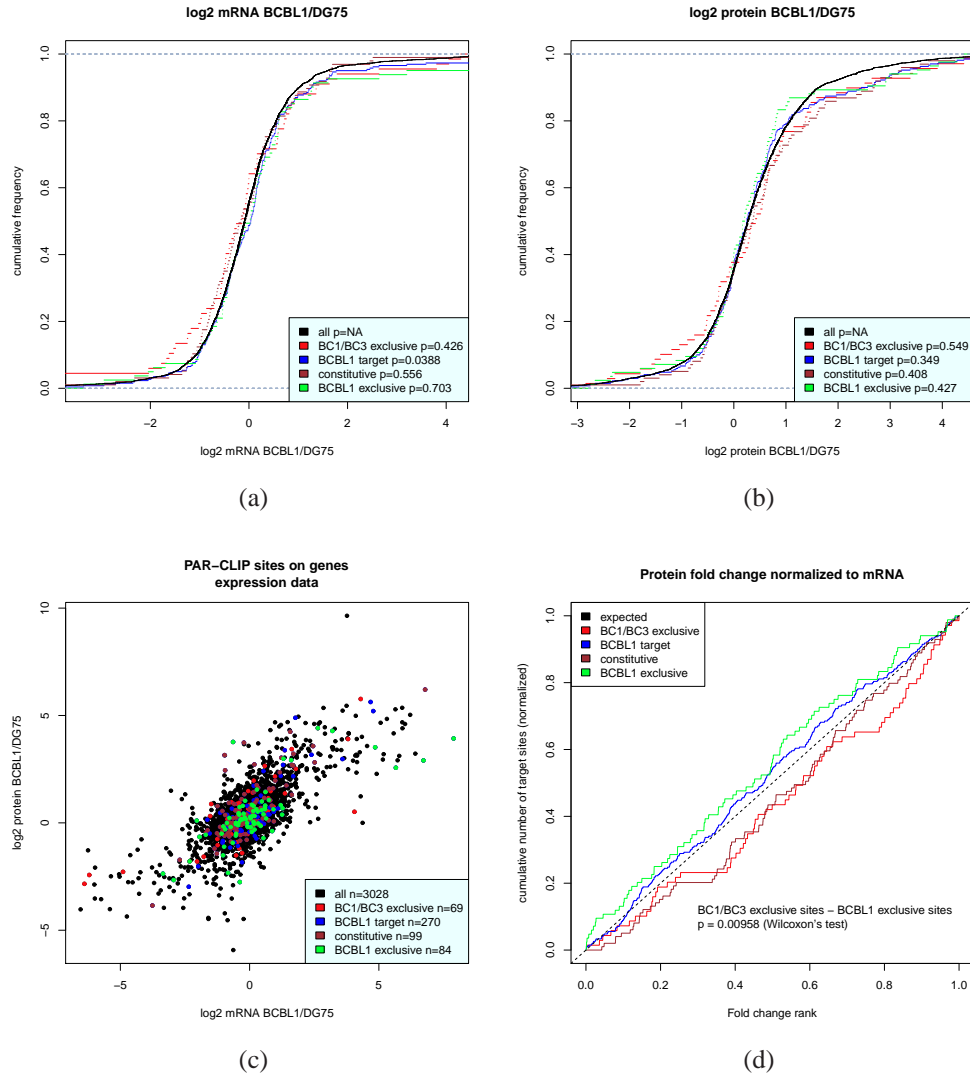


Figure 6.9: KSHV PAR-CLIP targets in expression data. Figures 6.9a and 6.9b show the fold change distributions of mRNAs and proteins between BCBL1 and DG75, respectively. When the log fold changes of mRNAs and proteins are considered individually, no significant shift for any set of context-specific microRNA targets is observed. In Figure 6.9c, genes are scattered according to their mRNA log₂ fold changes between BCBL1 and DG75 on the x-axis and to their protein log₂ fold changes on the y-axis. Target sites active in BCBL1 appear to be shifted towards the bottom-right. These sites correspond to genes whose protein level fold change between BCBL1 and DG75 is lower than expected from the mRNA level. Figure 6.9d shows the ranks of protein fold changes normalized to their mRNA levels for all gene sets considered. Normalized protein fold changes are significantly lower for genes with BCBL1 specific target sites than for genes with target sites inactive in BCBL1 ($p < 0.01$, Wilcoxon rank sum test).

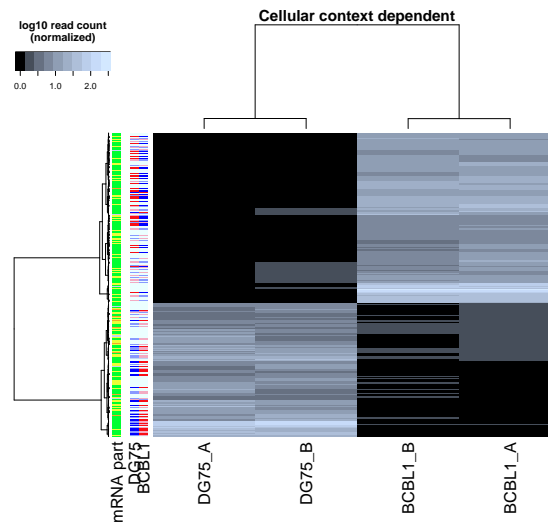


Figure 6.10: Context-dependent target interactions of human microRNAs. The differential PAR-CLIP analysis for all target sites of cellular microRNAs is visualized. The top part corresponds to BCBL1 specific target sites of constitutively expressed cellular microRNAs ($n = 184$) whereas the bottom half represents target sites exclusively active in DG75 and not in BCBL1 ($n = 137$). Importantly, all these patterns of context-dependency are highly reproducible across replicates.

Taken together, RIP-Chip data, RNA half-life data as well as mRNA and protein expression data provides good evidence that a substantial amount of KSHV microRNA target sites as found by differential analysis of PAR-CLIP data is indeed context-dependent which leads to a differential association with RISC and results in context-dependent functional impact on target gene expression.

6.3.4 Context-dependent target sites of cellular microRNAs

We next selected context-dependent microRNA/target interactions that are either active in BCBL1 or DG75 but not in both. Thus, we first selected all microRNAs that are not differentially expressed between BCBL1 and DG75 (< 2 -fold) and are reliably detected in the PAR-CLIP experiments (at least 100 reads in all four datasets). Furthermore, all microRNAs had to have at least 20 target sites as identified by a 7-mer seed by PARma. All identified microRNAs showed a clear pattern of context-dependency in their target sites. Using the same criteria as in the analysis of KSHV microRNAs, context-dependent target sites were defined (Figure 6.10).

Again, context-dependent microRNA/target interactions as defined by the differential PAR-CLIP analysis resulted in highly significant differential association with RISC (Figure 6.11a). Specifically, context-dependent targets are more than 2-fold enriched in significantly differentially RISC-associated mRNAs (PC2 score > 2) for both cellular contexts. Furthermore, target mRNA half-lives are again significantly lowered by the context-dependent activity of the microRNA/target interactions ($p < 0.0002$, Wilcoxon rank sum test; Figure 6.11b). Thus, as in

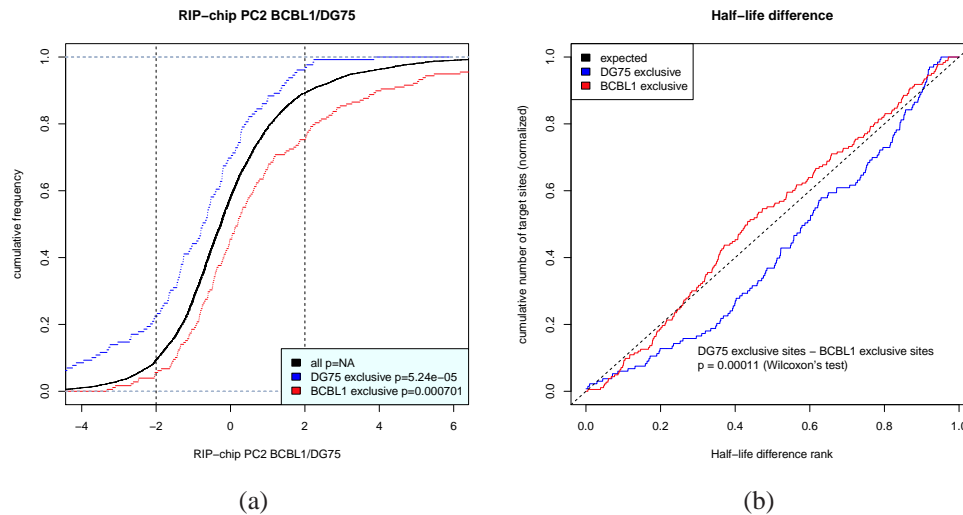


Figure 6.11: Cellular PAR-CLIP targets in RIP-Chip and mRNA half-life data. Figure 6.11a shows the distributions of the differential RIP-Chip scores as compared to all genes with any PAR-CLIP target sites (see also Figure 6.7a). Both, targets exclusively active in DG75 as well as in BCBL1 are significantly shifted towards stronger association with RISC in their respective context. The vertical lines indicate a threshold for strongly differentially RISC-associated genes. In both cases, the respective context-dependent targets are more than 2-fold enriched over the background genes (about 10% of background genes in comparison to > 20% of the target genes in both cases). In Figure 6.11b, the rank distribution of half-life differences for both sets of context-dependent targets is shown (see also Figure 6.8b). BCBL1 specific targets are significantly shifted towards lower half-life difference ranks in comparison to DG75 specific targets indicative for effects of context-dependent microRNA/target interactions in the respective context only.

the analysis of KSHV microRNAs, context-dependent target sites of cellular microRNAs also lead to differential RISC-association and have functional impact on target mRNA half-lives in a context-dependent manner.

The analysis of steady-state expression levels revealed a clear pattern of context-dependent targets: Both sets of context-dependent targets are clearly shifted in comparison to the background with respect to both mRNA and protein fold changes (Figures 6.12a and 6.12b). Specifically, genes tend to have higher expression in the context where the microRNA/target interactions are active.

Importantly, this is not solely due to a completely abrogated expression in the non-active context, since proteins are detected for almost half of all context-dependent targets in both cell lines and in more than two thirds of the cases, the fold change is smaller than 2-fold (Figure 6.12c). Thus, it is not the absence or presence of target mRNAs that lead to context-dependency of target sites. Rather, this indicates a complex dependency of the target site activity on the exact target mRNA expression levels. However, there may be a subpopulation within both sets of context-dependent

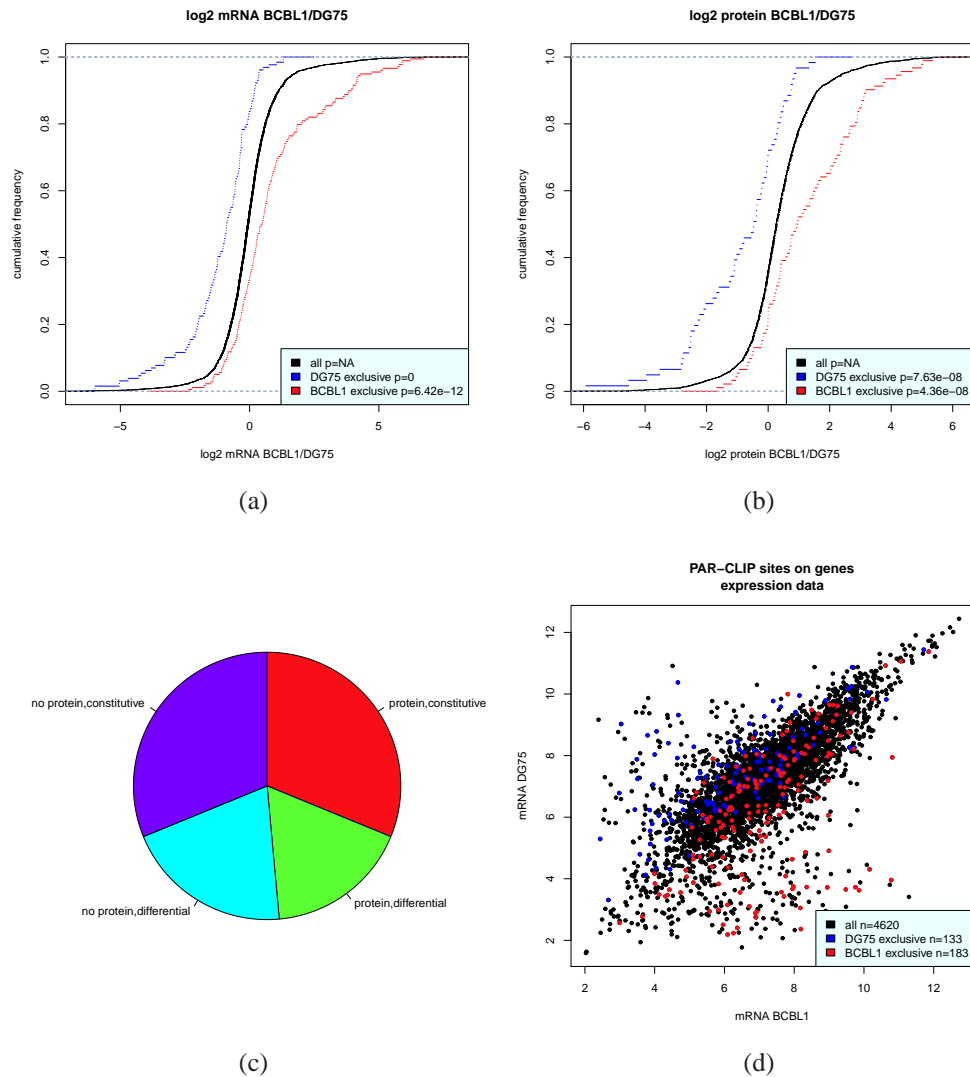


Figure 6.12: Cellular PAR-CLIP targets in expression data. Here the distributions of mRNA and protein fold changes between BCBL1 and DG75 for context-dependent targets of cellular microRNAs are shown, as compared to the background of all genes with any PAR-CLIP target site. Clearly, based on mRNA as well as on protein levels, context-dependent targets are higher expressed in their target context. This indicates that the target mRNA expression directly contributes to the cellular context of microRNA-mediated regulation. As depicted in Figure 6.12c, this is not solely due to a complete absence of gene expression in the non-target context, as proteins are detected for all these genes in half of the cases and more than two thirds are only slightly differentially regulated (< 2 -fold). Figure 6.12d shows a scatterplot of the microarray intensity measurements for all genes with a PAR-CLIP target site.

targets, where a missing activity of a target site may be explained by the complete absence of the target mRNA (Figure 6.12d).

Table 6.1: Identified motifs by MERCI. We searched for motifs in flanking sequences (\pm 80 bp) of context-dependent seed sites not explained by differential mRNA levels. These motif searches were done in a discriminative manner, i.e. by comparing a positive set to a negative set of sequences. E.g. BCBL1 exclusive sites of cellular microRNAs (Cellular BCBL1) were compared to DG75 exclusive target sites of cellular microRNAs (Cellular DG75). For each comparison, the identified motifs only occurred in the positive set and not in the negative set.

Positive set	with motif	Negative set	with motif	Min occurrences	Motif count
Cellular BCBL1	94/107	Cellular DG75	0/76	7	29
Cellular DG75	65/76	Cellular BCBL1	0/107	5	25
Viral BCBL1	83/100	Viral BC1/BC3	0/99	6	29
Viral BC1/BC3	74/99	Viral BCBL1	0/100	6	20

6.3.5 mRNA levels and flanking sequence motifs explain context-dependent microRNA/target interactions

Thus, we analyzed to which extent mRNA expression levels contribute to the cellular context and whether there are other factors that are necessary to explain the widespread context-dependency of target sites. First, we tested whether the target mRNA level is the only contributor that constitutes the cellular context for microRNA-mediated gene regulation.

Read counts are not only subject to biological variance but also to a substantial amount of sampling noise since many clusters only have a few dozen reads. To compare PAR-CLIP read count fold changes with mRNA fold changes in a more robust manner, it is therefore important to estimate the extent of this sampling noise. We used a population based estimate of variance using a conditional gamma distribution (see Methods). This approach is similar to recent methods to estimate significance of differential expression in RNA-seq data [Anders and Huber, 2010; Robinson et al., 2010].

If this noise model is applied to the comparison of mRNA fold change corrected PAR-CLIP target sites, more than 50% of all context-dependent target sites, i.e. at least 14% of all target sites of the selected set of cellular microRNAs, cannot be explained as judged by the P-value distribution (Figure 6.13). This means that in these cases, the PAR-CLIP read count fold change is significantly higher than expected from the corresponding mRNA fold change and this difference also cannot be explained by sampling noise inherent to low-count data such as PAR-CLIP. Thus, target site activities are not simply linearly dependent on mRNA levels.

Furthermore, as illustrated in Figure 6.13b, there are several instances where the target gene is not differentially expressed (i.e. datapoints around 0 on the mRNA \log_2 fold change axis) but where the target sites show a > 16 -fold elevated activity. In these cases, mRNA expression alone clearly cannot explain target site activity. Thus, other factors contribute to context-specific microRNA function.

RNA binding proteins (RBPs) likely constitute such additional contributors. Thus, we performed a motif search in regions flanking context-dependent target sites (seed site \pm 80 bp). For motif discovery we used MERCI [Vens et al., 2011], which is based on efficiently enumerating all

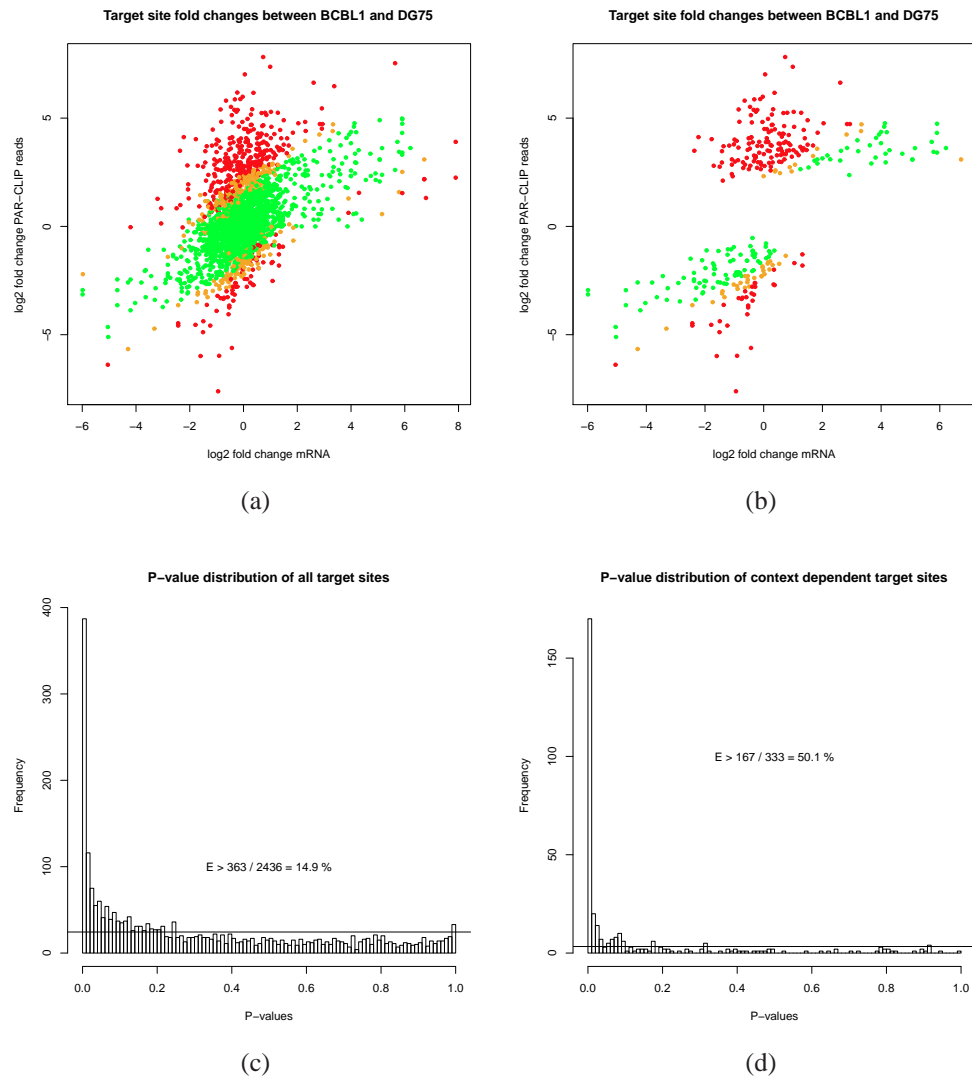


Figure 6.13: Comparison of mRNA fold changes to PAR-CLIP read count fold changes. Figure 6.13a shows a scatterplot comparing mRNA fold changes to PAR-CLIP read count fold changes of all target sites of the cellular microRNAs analyzed. For the PAR-CLIP data, a pseudocount of 1 was used. Green dots represent target sites that can be explained by the mRNA fold change while respecting sampling noise of the read counts, whereas orange and red dots correspond to significant outliers ($p < 0.05$ and $p < 0.01$, respectively). The P-value distribution in Figure 6.13c of all these target sites suggests that at least 14.9% (363 instances with $p < 0.01$ of overall 2436 target sites after subtraction of baseline indicated by the horizontal line) of all differential target site activities cannot be explained by the mRNA fold change and sampling noise. Figures 6.13b and 6.13d illustrate this for the context-dependent microRNA/target interactions only. Here, more than 50% of all sites cannot be explained by mRNA levels.

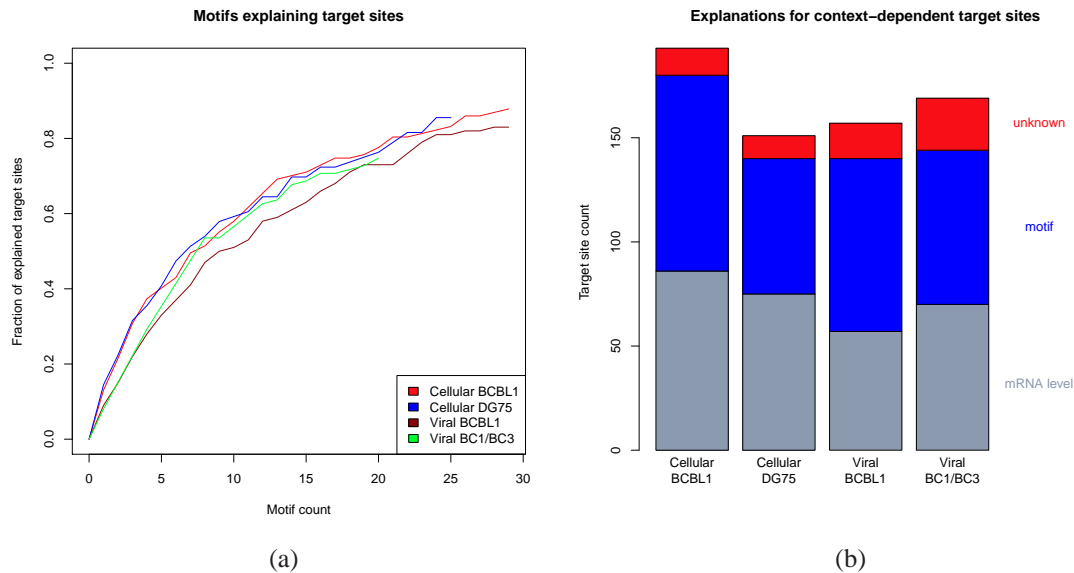


Figure 6.14: Role of sequence motifs for context-dependent target sites. Figure 6.14a shows the fraction of context-dependent target sites that contain a certain number of discriminative k-mers. Only target sites that cannot be explained by mRNA levels were used. A k-mer is discriminative if it occurs n times in the positive set (e.g. cellular BCBL1 exclusive sites in red) and does not occur in the corresponding negative set (e.g. cellular DG75 exclusive sites, see Table 6.1). We sorted discriminative k-mers according to their number of occurrences in decreasing order and chose a cutoff for n based on our randomization experiments (Figure 6.15). In all cases, between 75% and 90% of all context-dependent target sites can be explained by a discriminative k-mer. In Figure 6.14b, putative explanations for the full sets of context-dependent target sites are illustrated. On average, more than 90% can be explained by either differential mRNA levels or the presence of a discriminative k-mer.

discriminative k-mers of two sets of sequences. Specifically, we searched for k-mers that do not occur in the negative set and occur at least n times in the positive set and we only considered target sites from mRNAs that are not differentially expressed. n was chosen according to the total number of sequences in the positive set. MERCI identified 20-30 k-mers when we compared target sites of cellular microRNAs exclusively present in BCBL1 to those exclusively present in DG75 and target sites of viral microRNAs exclusively present in BCBL1 to those in BC1/BC3 or vice versa (Table 6.1 and Figure 6.14a). These discriminative k-mers occur in 75%-90% percent of all context-dependent target sites that cannot be explained by the mRNA level and as few as 5 motifs already can explain 30%-40% of all sites. In contrast, discriminative k-mers found by chance in randomized sequences only occur in a considerably lower number of sequences (Figure 6.15). Thus, these motifs are likely candidates of binding sites for RBPs contributing to context-dependent recognition of target sites by microRNAs.

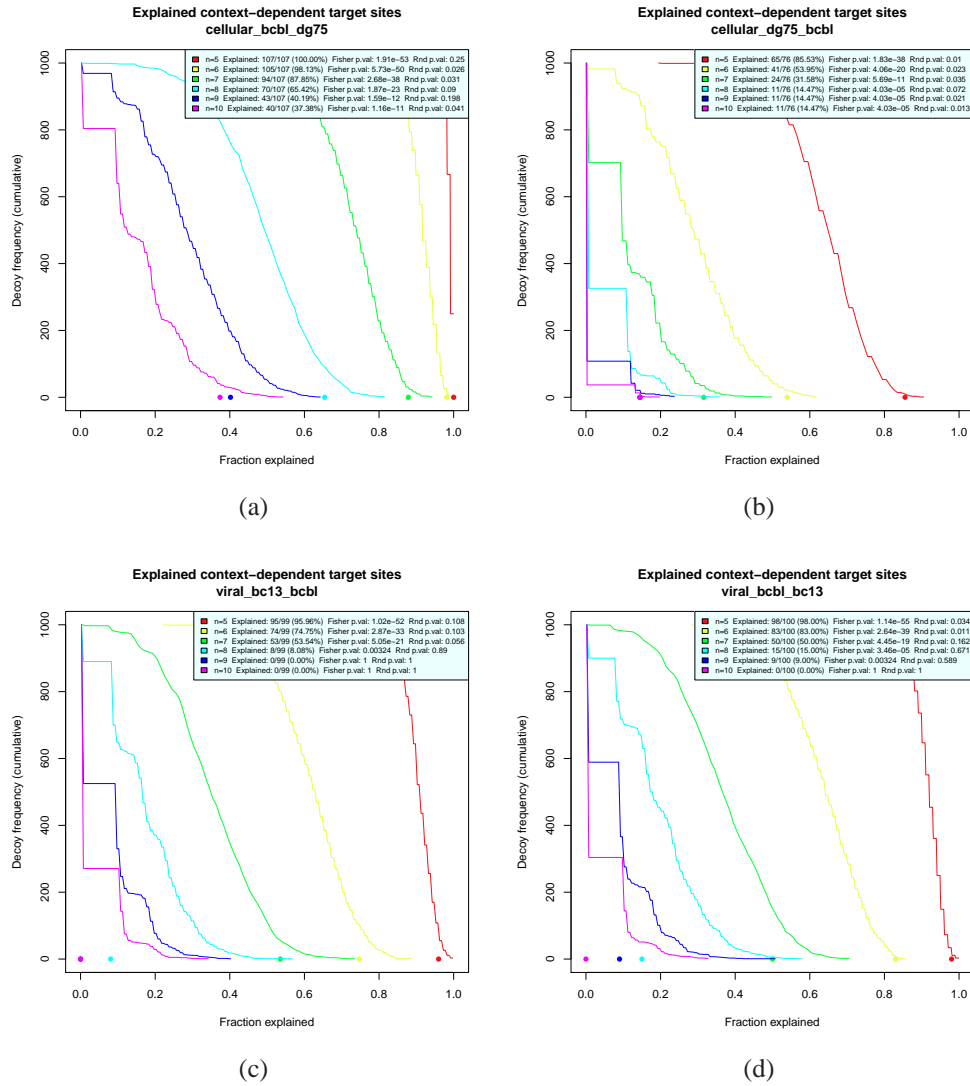


Figure 6.15: Motif randomization results. For each comparison, positive and negative labels were randomly permuted 1000 times. MERCI was run on each randomized instance and the number of sequences containing an n -discriminative k -mer was counted for n between 5 and 10. A k -mer is n -discriminative, if it occurs in at least n sequences in the positive set and does not occur in the negative set. We plotted the distributions of the fractions of explained sequences and compared them to the actual fractions in the true positive and negative sets (points in the plots).

In summary, from all context-dependent target sites identified by PAR-CLIP and validated by RIP-Chip experiments, 4sU tagging based mRNA half-lives and mRNA and protein expression measurements, more than 90% can either be explained by differential mRNA levels or by the presence of a putative RBP binding motif (Figure 6.14b).

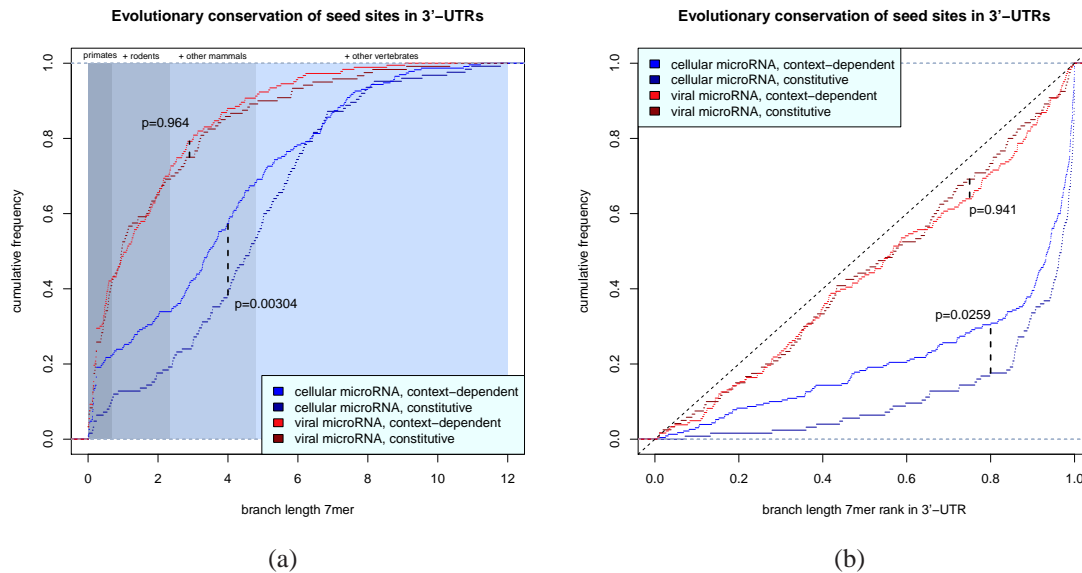


Figure 6.16: Conservation of target sites. Distributions of branch lengths of target sites are illustrated in Figure 6.16a (see main text for a definition of branch lengths). Shaded regions indicate the maximal branch lengths of target sites conserved in primates, in primates and rodents, in mammals and in vertebrates. All cellular microRNAs considered here are conserved in vertebrates. Constitutive target sites of these microRNAs are significantly more conserved ($p < 0.00304$, two-sided Kolmogorov-Smirnov test) than context-dependent target sites. Moreover, neither context-dependent nor constitutive target sites of viral microRNAs show evidence for evolutionary conservation. Figure 6.16b shows that these patterns are not due to different overall 3'-UTR conservation levels of target mRNAs. Branch lengths were computed for all 7-mers in each 3'-UTR and the distribution of the rank of the seed (normalized between 0 and 1) among all corresponding 3'-UTR branch lengths was considered.

6.3.6 Context-dependent target sites are less conserved than constitutive sites

Finally, we asked whether context-dependent target sites have distinct evolutionary conservation patterns as compared to constitutive target sites. Following the approach of Friedman et al. [2009], for each target site we computed the branch length along the phylogenetic tree of 46 vertebrates by summing all branches where the seed of a cluster is fully conserved in the genome-wide multiple alignment of 46 vertebrate species. The branch length thus incorporates both the evolutionary age as well as the loss of a target site in specific lineages. Specifically, a target site that emerged in the last common ancestor of primates and rodents, and has not been lost in any primate or rodent lineage has a branch length of 2.342 (shaded areas in Figure 6.16).

Intriguingly, constitutive target sites of conserved cellular microRNAs are significantly stronger conserved than context-dependent sites ($p < 0.003$, two-sided Kolmogorov-Smirnov test). For

instance, while more than 80% of constitutive sites are conserved beyond the last common ancestor of primates and rodents, only about 65% of context-dependent sites are conserved beyond this clade. 20% of context-dependent sites even show a signature of recent evolution within the primate lineage. Importantly, this does not reflect the overall conservation level of the respective 3'-UTRs, but is specific to the seed sites (Figure 6.16b).

Target sites of viral microRNAs, independent whether they are context-dependent or constitutive, show patterns of much weaker conservation. This can be expected, as there are no conserved viral microRNAs and as pathogenicity of KSHV may rather induce positive selection of its microRNA target sites on host mRNAs.

6.4 Discussion

In this study, we analyzed PAR-CLIP data from four human B-cell lines, three of which are infected with Kaposi's sarcoma-associated herpesvirus (KSHV), using an improved computational approach to identify target sites of both cellular and viral microRNAs (PARma, Erhard et al. [2013a]). The overlap in target sites between the four cell lines was surprisingly low (about 40%), indicating a large set of context-dependent microRNA/target interactions. Three additional sets of high-throughput data (RIP-Chip, 4sU-tagging-derived RNA half-lives and SILAC proteomics data) supported this observation: Context-dependent microRNA targets are associated with RISC in a context-dependent manner and have a measurable functional impact on their targets in a context-dependent manner. This was observed for the targets of both, cellular and viral microRNAs. The latter offered an important control as they were exclusively observed in the cells expressing the viral microRNAs. Thus, we propose a new layer of complexity in microRNA targeting: Depending on the cellular context, specific microRNA/target interactions may be active or not, even if both microRNA and target mRNA are expressed. Furthermore, we could show that the evolutionary conservation differs between context-dependent and constitutive target sites, indicating that selective pressure may be different for context-dependent and constitutive target sites or that they have different evolutionary ages.

6.4.1 Contributors to the cellular context

Cellular context may be formed directly by the quantities of microRNAs and mRNAs: Dependent on the exact copy numbers of microRNAs and mRNAs in each cell, intricate regulatory mechanisms may emerge leading to highly complex patterns of regulation [Mukherji et al., 2011]. Furthermore, due to the many-to-many relationship of regulators and targets, microRNAs and mRNAs are embedded in a highly complex regulatory network [Hobert, 2008]. Our analyses indicate that the quantities of microRNAs and target mRNA are direct contributors to the cellular context. However, based on our results, more than 50% of all observed context-dependent microRNA/target interactions cannot be explained by microRNA or mRNA levels and, therefore, are likely dependent on indirect factors.

The presence of RNA binding proteins (RBPs) may prevent microRNA binding to nearby sites [Bhattacharyya et al., 2006] or also induce binding [Kim et al., 2009a]. In a recent study the

whole RNA binding proteome of a cell line was examined by PAR-CLIP coupled to high resolution mass spectrometry [Baltz et al., 2012]. This study revealed two important aspects of RBPs: First, in a single cell type, about 800 different RBPs can be identified. This unexpectedly high number of RBPs allows for highly complex combinatorics of competitive or activating RBP-microRNA interactions. And second, crosslinking events were observed for almost 30% of all uridines in 3'-UTRs, suggesting that mRNAs are broadly covered by RBPs. Indeed, we could identify a handful of sequence motifs that are able to explain a large fraction of context-dependent target sites, indicating that RBPs may play important roles in shaping the cellular context for microRNA-mediated regulation.

Thus, there is an intriguing analogy of the transcriptional and post-transcriptional layer of regulation: DNA, which is the material for transcriptional regulation, is covered by histones, transcription factors and other DNA binding proteins and the composition and dynamics of these proteins contribute to the cellular context [Consortium, 2012b]. This cellular context determines to which extent a certain transcription factor can bind to a specific target site and exert its regulatory role. Context-dependent regulatory networks may differ dramatically across different cell types or conditions [Neph et al., 2012a]. Similarly, mRNAs, which are the units for post-transcriptional regulation, are covered by RBPs, and we argue that their composition and dynamics contribute to a cellular context for microRNA-mediated regulation. Additionally, factors other than these covering proteins may further shape the cellular context for both, transcriptional and post-transcriptional regulation: For transcriptional regulation, distinct modifications of chromatin or the DNA may also determine context. Furthermore, chromosomal conformations may place distal binding sites of transcription factors to promoters of different genes in three-dimensional space and may therefore also be important.

6.4.2 Other contributors

mRNAs may even provide more opportunities for context-dependent regulation: While DNA usually is restricted to a single cellular compartment, the nucleus, the life cycle of mRNAs may span multiple compartments and subcompartments. This cellular localization may itself be regulated and depending on the localization, mRNAs may be translated or not. For instance, sequestering of mRNAs to P-bodies by microRNAs leads to a reduced translation and mRNA decay [Pasquinelli, 2012]. Furthermore, the single stranded mRNA gives rise to complex secondary and tertiary structures, and it has been shown that the accessibility of target sites determines whether microRNAs can bind to the mRNA or not [Kertesz et al., 2007]. Interestingly, the conformation of RNAs is highly flexible and may be reshaped in a context-dependent way: Kedde et al. [2010] have shown that the activation of the RNA binding protein Pumilio-1 induces a local change in a hairpin structure of the 3'-UTR of the p27 tumour suppressor mRNA. Upon Pumilio-1 activation, an inaccessible binding site of miR-221/miR-222 is opened for binding, leading to an efficient repression of p27.

In addition, RISC is a highly modular protein complex [Frohn et al., 2012]. Therefore, proteins that interact with RISC may influence the effects of microRNA/target interactions: For instance, the NHL family protein LIN41 has been found to suppress let-7 and miR-124 activity by ubiquitilation of AGO2 [Rybak et al., 2009] and several other NHL proteins have been implicated

in the regulation of RISC activity [Pasquinelli, 2012]. In addition to ubiquitilation, AGO2 is susceptible for several other types of modification including hydroxylation [Qi et al., 2008], phosphorylation [Rüdel et al., 2011] and poly(ADP)-ribosylation [Leung et al., 2011].

Another layer of complexity in microRNA-mediated regulation is induced by mutual microRNA-target regulation. Very strong microRNA binding sites in an mRNA [Franco-Zorrilla et al., 2007], a pseudogene [Cazalla et al., 2010], a non-coding RNA [Cesana et al., 2011] or viral RNAs [Marcinowski et al., 2012] may sequester RISCs containing a specific microRNA acting as a microRNA sponge. It has also been hypothesized that the microRNA target sites in a cell as a whole allow for crosstalk between expressed transcripts giving rise to an intricate post-transcriptional regulatory network based on mutually competing microRNA/target interactions [Salmena et al., 2011]. The complexity of such a regulatory network is further underlined by the nonlinearity of the regulatory outcome of a microRNA/target interaction [Mukherji et al., 2011]: Depending on the exact copy numbers of microRNA and mRNA and the affinity of the microRNA for the target site, protein expression may be either completely abolished or only fine-tuned.

It has been suggested that sequestering of RISCs by the expression of transcripts containing one or multiple strong target sites for a specific microRNA may derepress its targets [Franco-Zorrilla et al., 2007; Cazalla et al., 2010; Cesana et al., 2011; Marcinowski et al., 2012]. Thus, such microRNA sponges may also be important contributors to the cellular context for microRNA-mediated regulation. In such a setting, weak target sites should disappear first. By comparing binding energies for our set of context-dependent target sites, we tested whether such effects play a role in our datasets. However, we could not identify any microRNA where exclusive binding sites had significantly different binding energies than constitutive target sites (data not shown). This may be due to deficiencies of the current RNA energy model to describe microRNA/target duplexes, or because microRNA sponges do not play an important role for our cell lines. And indeed, all of the selected cellular microRNAs exhibit target sites that are exclusive in DG75 and other target sites exclusively present in BCBL1, which would not be expected if a microRNA sponge is active in one of these cell lines.

6.4.3 Functional considerations of context-dependent regulation

Based on results from concurrent research on transcriptional regulation [Consortium, 2012b; Wang et al., 2012b,a; Yanez-Cuna et al., 2012], context-dependency in post-transcriptional regulation should not come as a surprise: It is known that transcription factors bind to their target sites in a context-dependent manner. Therefore, context-dependency of regulatory mechanisms presumably is beneficial in an evolutionary sense, and this is a widespread phenomenon for transcriptional regulation. Here, we argue that evolution also has invented this additional layer of complexity for microRNA-mediated regulation as well.

One evolutionary benefit of the additional layer of complexity by context-dependent microRNA/-target interaction may be the greater flexibility in regulation: Modulating the expression level of a microRNA would alter the regulation of hundreds of targets and therefore potentially influence a multitude of cellular processes. In contrast, using context-dependent regulation, for instance by activating or inactivating an RNA binding protein (RBP), smaller groups of targets could be

activated or inactivated in a much more focused manner. The combinatorics that unfolds when multiple RBPs, multiple target sites or other factors contribute to the overall regulation provides opportunities for evolutionary forces to achieve the desired expression levels for individual genes. Our analysis of target sites of constitutively expressed cellular microRNAs revealed that a large fraction of context-dependent targets may be due to induced mRNA levels. For instance, a gene may get transcribed at high rates in BCBL1 as compared to DG75, leading to elevated mRNA levels. At the same time, microRNA target sites are more active in BCBL1 than in DG75, leading to an induced degradation as compared to DG75. Importantly, this dependency between microRNA and mRNA is not necessarily linear, as pointed out above. Thus, these constitutive microRNAs seem to limit the expression levels of their target mRNAs: If targets have high enough expression levels, they become subject to microRNA-mediated regulation thus providing an upper bound for the target mRNA levels.

6.4.4 Consequences of context-dependency

The differential analysis of a collection of high-quality large-scale experiments for microRNA target site discovery indicates that context-dependent microRNA targeting is not restricted to a few examples, but is a widespread phenomenon and a general feature of microRNA mediated regulation. This has significant consequences for both computational and experimental approaches for microRNA target discovery.

MicroRNA target prediction algorithms may not as bad as their reputation [Thomas et al., 2010; Sethupathy et al., 2006; Ritchie et al., 2009]: False positive as well as false negative predictions simply may be due to a wrong context used when evaluating the predictions. Thus, none of the apparently inconsistent evaluations of microRNA target prediction algorithms may be wrong: Each of these was evaluated on a different cellular context and, consequently, differing prediction methods seemed more accurate than others. However, the problem of microRNA target prediction may be defined in an incorrect way and as long as prediction methods do not incorporate the cellular context, predicted targets are of limited use. Thus, we expect that future development of microRNA target prediction methods will mainly depend on integrating features of cellular context into the prediction algorithms. Such an approach is obviously heavily dependent on progress in unraveling contributing factors to the cellular context.

Another consequence of a general context-dependency of microRNA targeting is that experimental assays for microRNA target discovery and validation must be interpreted with care. In various studies, either a single or a pool of microRNAs has been transfected into a cell line and gene expression has been measured genome-wide either on the mRNA level using microarrays or RNA-seq [Lim et al., 2005; Linsley et al., 2007; Grimson et al., 2007; He et al., 2007; Xu et al., 2010] or on the protein level using mass spectrometry [Selbach et al., 2008; Baek et al., 2008] differentially for transfected and control cells. In addition to the well-known problem of secondary regulation [Tu et al., 2009; Naeem et al., 2011], there are several reasons why downregulated genes should not generally be taken as the set of targets for the transfected microRNA: First, they may be targets exclusively active in the cell line investigated in the study. Second, the transfected microRNA may have copy numbers at levels never occurring in physiological conditions. And third, the transfection itself may lead to an altered cellular

context, e.g. by the induction of cellular stress pathways. Also, further conclusions drawn from such experiments, e.g. that microRNAs generally lead to widespread but only modest regulation should be revisited: This may only be true in the context of the experiment, which is not a cellular context that evolved naturally but has been forced onto a cell line artificially. Furthermore, later studies revealed that the particular outcome or strength of regulation is dependent on the exact mRNA and microRNA copy numbers [Mukherji et al., 2011].

Also, the most widely used validation assay for microRNA targets is based on fluorescence reporter genes that are fused to target 3'-UTRs and co-transfected with the microRNA into a particular cell line usually lacking this microRNA [Kiriakidou et al., 2004; Gottwein et al., 2011]. Obviously, not only is the microRNA-target pair introduced into a non-natural context and probably expressed at non-physiological conditions, but also the microRNA target site itself is expressed in a non-natural context, i.e. the 3'-UTR of a fusion gene. Thus, both outcomes of such an experiment may not hold for the microRNA target pair under different conditions, i.e. if the reporter assay does not confirm regulation, the target site may still be highly functional in the right context and if the reporter assay validates regulation, this may not be true for the context under consideration. In conclusion, while luciferase assays provide good evidence that a certain gene can be regulated by a given microRNA it does not allow any claims about whether this interaction is of relevance in the biological context of interest.

Consortium-driven endeavors to unravel context-dependent transcriptional regulation have been started years ago, as soon as the human genome project has been finished [Consortium, 2012b]. Context-dependent transcription factor binding sites have been determined in an overwhelming variety of conditions and it is one of the most intriguing results of the ENCODE project that context-dependency is one of the key features of transcriptional regulation.

Here, we show that context-dependency is also an important factor in post-transcriptional regulation. We propose that a similar approach as for transcriptional regulation must also be taken for microRNA-mediated regulation and that a lot of additional experiments are necessary to further investigate both, microRNA targets specific to certain contexts and key contributors that determine cellular context.

6.5 Methods

6.5.1 Cell lines

DG75-eGFP and BCBL1 were cultured in RPMI medium supplemented with 10% fetal calf serum and pen/strep.

6.5.2 PAR-CLIP and sequencing

PAR-CLIP on DG75 and BCBL1 were performed by the Zavolan laboratory as described [Kishore et al., 2011; Jaskiewicz et al., 2012]. Briefly, a total of 3×10^8 cells per replicate were grown and treated with 4-thiouridine (Sigma) for 14 hours (final concentration 100 μ M). Cells were pelleted and washed in cold PBS. Aliquots of 5×10^7 cells were resuspended in 5 ml of

cold PBS, placed in a 15 cm petri dish and irradiated at 365 nm with 100 mJ twice on ice, with 30 s break in between. Cross-linked cells were collected, pelleted and snap-frozen. PAR-CLIP was performed using 11A9 anti-Ago2 monoclonal antibody [Rüdel et al., 2008]. The PAR-CLIP sequencing data for BC1 and BC3 from [Gottwein et al., 2011] have been downloaded from GEO (accession number: GSE32113). We applied PARma [Erhard et al., 2013a] to the whole collection of all PAR-CLIP datasets as described.

6.5.3 SILAC-based proteomics

SILAC and LC-MS/MS were performed as described in the Mann laboratory at MPI for Biochemistry in Munich. The raw files from the mass spectrometer have been analyzed using MaxQuant (version 1.2.2.5) [Cox and Mann, 2008] using standard parameters against all human proteins from Ensembl (v60).

6.5.4 RIP-Chip analysis

For the RISC-IPs, 5×10^8 cells were taken for each replicate and processed as previously described [Dölken et al., 2010] using 6 μ g of purified monoclonal hAgo2 antibody (α -hAgo2; 11A9) or monoclonal BrdU-antibody (Abcam; used as control).

6.5.5 RNA half-life measurements by 4sU-tagging

The RNA half-life data for DG75 and BCBL1 have been published previously [Dölken et al., 2010]. In brief, newly transcribed RNA was labeled for 1h by adding 100 μ M 4sU to the cell culture medium. Total RNA was prepared using Trizol and newly transcribed RNA was purified as described [Dölken et al., 2008]. Three replicates of newly transcribed, total and preexisting RNA were measured.

6.5.6 PARma

PARma is specifically designed to accurately determine target sites and to determine which microRNA is responsible for each target site and is described in a separate paper [Erhard et al., 2013a]. Briefly, it estimates seed activity probabilities and the parameters of a generative model for the PAR-CLIP data simultaneously in an iterative manner. The model and probabilities are then used to accurately determine the seed position within each PAR-CLIP cluster, and to compute a cluster confidence score (C-score) and a microRNA assignment confidence score (MA-score). The C-score can be used to exclude false positive clusters (i.e. clusters that do not correspond to a target site of any microRNA), whereas the MA-score can be used to judge whether the assigned microRNA is indeed targeting a given site.

The PAR-CLIP expression value for each cluster is computed for each experiment by counting the reads overlapping the main crosslinking site [Erhard et al., 2013a]. For proper comparison across experiments, the expression values for all clusters are normalized using the same strategy as in Anders and Huber [2010], i.e. by dividing each count value by the geometric mean

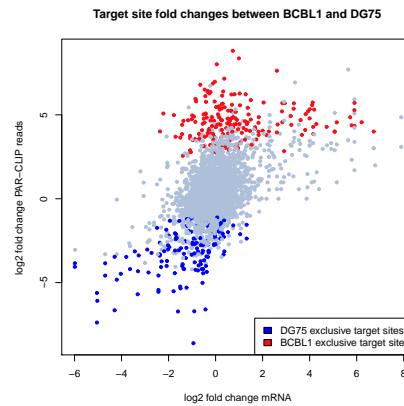


Figure 6.17: PAR-CLIP read count correlation with expression; The mRNA expression fold change is scattered against the PAR-CLIP read count fold change. Replicate counts were summed and a pseudocount of 1 was used to circumvent division by zero. In red and blue, context-dependent target sites of cellular microRNAs are shown.

across experiment and taking the median of all these values from a specific experiment as the normalization factor for this experiment.

To analyze the positional distribution, we subdivided each transcript into 60 bins and counted the number of target sites of cellular, EBV and KSHV microRNA, respectively, belonging to each bin. In order to compare cellular and viral frequencies, the number of target sites within each bin was divided by the total number of cellular, EBV or KSHV target sites, respectively.

Context-dependent target sites were determined by applying stringent cutoffs: More than 10 normalized reads in all replicates in the active context and less than 5 in all other experiments.

6.5.7 Correcting for sampling noise

In order to estimate the contribution of mRNA levels to the cellular context for microRNA-mediated regulation, we first inspected the correlation between the mRNA fold changes and PAR-CLIP read fold changes (see Figure 6.17). Replicate counts were summed and a pseudocount of 1 was used to circumvent division by zero. These fold changes were correlated to some extent, but there were also many exclusive (i.e. context-dependent) target sites present, that did not show any or only a very modest mRNA fold change. In order to properly estimate the fraction of non-correlated target sites and to handle the sampling noise of the low-count data and pseudocounts, we took the following approach:

First, we estimated the variance of PAR-CLIP fold changes based on replicate experiments. Because the number of replicates was extremely low ($n = 2$), no reliable estimates can be compute in a target site-wise manner, and, thus, we took a population based approach similar to methods that estimate significance of differential expression in RNA-seq data [Anders and Huber, 2010; Robinson et al., 2010]. Since the variance is not equal for strong target sites and weak target site (as measured by the number of PAR-CLIP reads) due to sampling noise, variance

was estimated conditional on the target site strength. Then, for each target site, we checked, whether the mRNA fold change was within a critical region as defined by significance levels of 1% and 5%.

To estimate read count fold change variances, we considered the absolute difference of read counts from replicate measurements (see Figure 6.18). For a given target site strength (as measured by the geometric mean of read counts across replicates), the distribution of these absolute differences resembled a gamma distribution by visual inspection. Thus, by using a running window approach, we estimated the distribution of absolute read count differences by fitting a gamma distribution to each window of 1000 target sites along the target site strength (i.e. the red line in Figure 6.18) using the *fitdistr* function from the R package MASS. We plotted the rate and shape parameters of the gamma distribution as fitted for different windows along the target site strength (see Figure 6.18c) and noticed that the shape parameter was relatively constant and the rate parameter increased linearly in log space with the target site strength. For robustness of the fits, we therefore computed the median shape parameter S across all windows and computed a robust linear fit for the rate parameters $R(s)$ against the logarithmized target sites strengths s . Thus, our model describes the absolute read count difference of a target site with strength s (i.e. the geometric mean of read counts across all experiments) by a gamma distribution with rate and shape parameters $R(s)$ and S .

This conditional gamma distribution allows us to compute the distribution of absolute read count differences for a given target site strength. As illustrated in Figure 6.18d, this conditional gamma distribution nicely reflects the variances for replicate measurements.

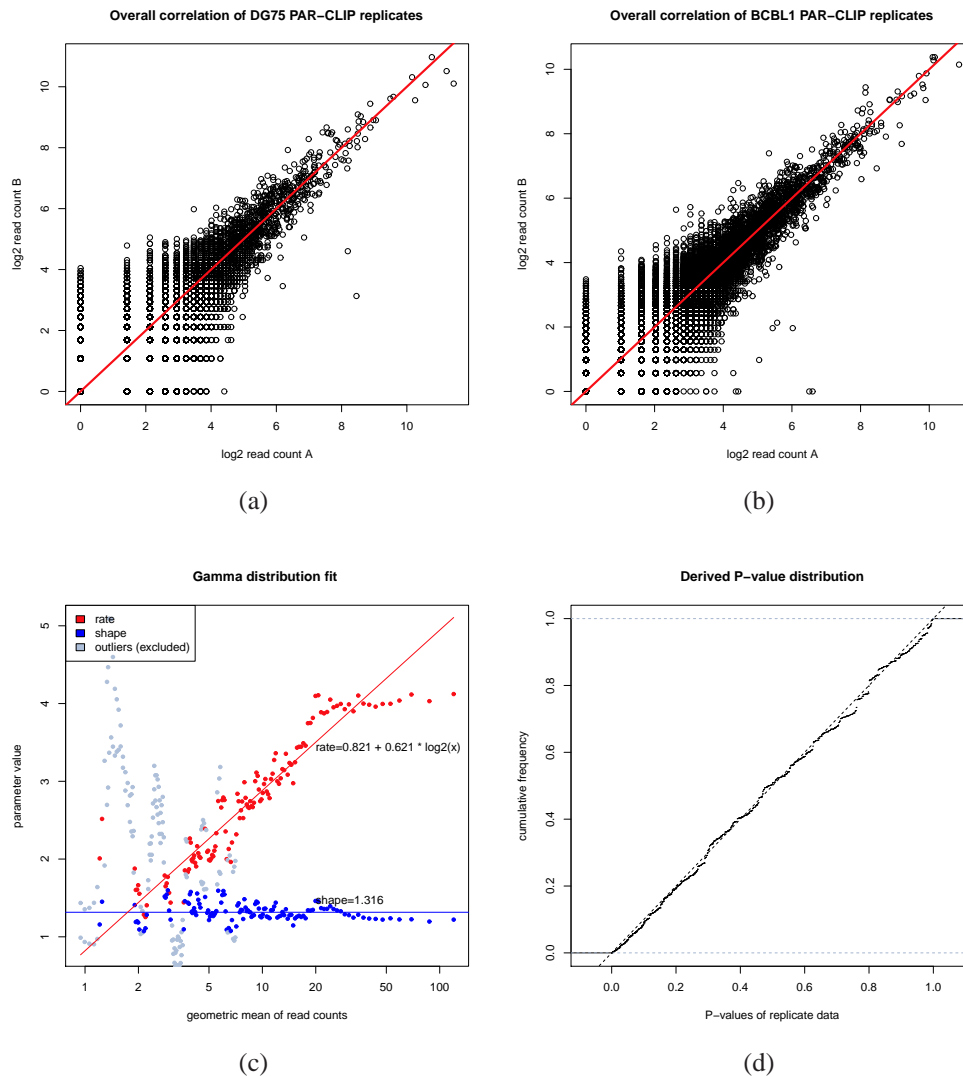


Figure 6.18: Conditional gamma distribution fit. In Figures 6.18a and 6.18b, the correlation of replicate PAR-CLIP read counts are shown in log scale for the DG75 and BCBL1 experiments, respectively (using a pseudocount of 1). The red line indicates the main diagonal. Deviations from the diagonal are obviously larger for weak targets sites (bottom left), indicative for sampling noise inherent to low count data such as PAR-CLIP. Figure 6.18c shows the fitted gamma distribution parameters against the target site strength. Outliers from the conditional model are indicated in gray (see Methods). In Figure 6.18d the p-value distribution of the gamma model applied to replicate measurements is shown. It closely resembles a uniform distribution, which indicates that our model accurately resembles the observed deviations in replicate measurements.

Chapter 7

Detection of outlier peptides

Motivation: While the previous chapters mainly focussed on viral microRNAs, their classification and targets, this chapter concentrates on possible effect of viruses and potentially of microRNAs, their impact on splicing patterns. To investigate alternative splicing patterns on large-scale, several methods have been proposed for RNA-seq data [Richard et al., 2010; Trapnell et al., 2013]. However, the effect of splicing only plays a role on protein level and considering the mRNA is only a proxy to proteins. Thus, I investigated whether and to which extent shotgun mass spectrometry data can be used for the identification of splicing patterns. In particular, I focussed on differential splicing, i.e. on splicing patterns that are different between two conditions. In order to develop a method to identify differential splicing and for its evaluation, I considered a publicly available high-quality dataset which I preprocessed using MaxQuant, which is a recent and widely used analysis software for raw LC-MS/MS data [Cox and Mann, 2008]. This is described in this chapter. I also applied this methods to the mass spectrometry data generated by our collaboration partners, which, however, did not yield any promising candidates for further experiments (see section 2.2.1).

Publication: An abstract of this chapter has been presented and published at the German Conference on Bioinformatics (GCB) 2011 in Weihenstephan, Germany [Erhard and Zimmer, 2011]. Subsequently, a full paper has been published in the Journal of Proteomics [Erhard and Zimmer, 2012]. Here, I adapted the layout and made minor corrections to the text.

My contribution: I came up with the method and the evaluations, implemented the method, carried out evaluations and wrote the paper.

Contribution of co-authors: Ralf Zimmer supervised the work and helped to revise the manuscript

7.1 Abstract

Quantitative high-throughput mass spectrometry has become an established tool to measure relative gene expression proteome-wide. The output of such an experiment usually consists of a list of expression ratios (fold changes) for several thousand proteins between two conditions. However, we observed that individual peptide fold changes may show a significantly different behavior than other peptides from the same protein and that these differences cannot be explained by imprecise measurements.

Such outlier peptides can be the consequence of several technical (misidentifications, misquantifications) or biological (post-translational modifications, differential regulation of isoforms) reasons. We developed a method to detect outlier peptides in mass spectrometry data which is able to delineate imprecise measurements from real outlier peptides with high accuracy when the true difference is as small as 1.4 fold.

We applied our method to experimental data and investigated the different technical and biological effects that result in outlier peptides. Our method will assist future research to reduce technical bias and can help to identify genes with differentially regulated protein isoforms in high throughput mass spectrometry data.

7.2 Introduction

Mass spectrometry (MS) based proteomics has become a common tool for a wide range of biological research areas [Baek et al., 2008; Cox and Mann, 2007; Huttlin et al., 2010; Schwanhauser et al., 2011; Selbach et al., 2008]. In a shotgun experiment, proteins from a complex sample are digested into peptides (e.g. using Trypsin) whose mass-to-charge ratios are then measured in a first round of MS after ionization. Metabolically (e.g. SILAC) or chemically (e.g. iCAT) introduced heavy amino acids can be used as labels to distinguish peptides in a mixture of samples in the same MS run [Ong and Mann, 2005]. Measurement intensities are related to peptide abundances and can therefore be used for quantification. These MS spectra alone do not provide a reliable way to identify peptide sequences in a complex sample since mass alone is not a reliable discriminator for peptides [Colinge and Bennett, 2007]. Therefore, tandem mass spectrometers are able to select one or several peaks per MS scan for further fragmentation followed by a second round of MS (MS^2 spectra). The most abundant fragments produced are so-called b and y ions, which are the result of fragmentation between the amino and hydroxy groups of two consecutive amino acids and correspond thus to prefixes and suffixes of the original peptide. It has been shown that these MS^2 spectra provide enough information to identify peptide sequences.

Primary data analysis is usually done by integrated analysis pipelines, e.g. TPP [Keller et al., 2005], TOPP [Bertsch et al., 2011] or MaxQuant [Cox and Mann, 2008]. In modern high-resolution LC-MS/MS settings, data analysis generally consists of the two crucial steps peptide identification and quantification.

For peptide identification, experimental MS^2 spectra are compared to theoretically computed spectra from peptides derived from a protein sequence database. Several methods to score

experimental to theoretical spectra have been developed and are available either as commercial software such as Mascot [Perkins et al., 1999] or Sequest [Yates et al., 1995] or as open source tools such as X!Tandem [Craig and Beavis, 2004] or Andromeda [Cox et al., 2011]. Such methods typically report a candidate list of possible sequences for each MS² spectrum with one or several associated scores. False discovery rates (FDR) can be calculated using a decoy database approach: For each protein in the database, a (pseudo-) reversed protein is created and also used for database search. For a given score cutoff, the FDR then is equal to the fraction of decoy identifications above this cutoff [Gupta and Pevzner, 2009; Käll et al., 2008].

Generally, there are two types of quantification: For an absolute quantification, the concentrations of all proteins within a single sample must be determined, whereas for relative quantification the concentration ratio (the fold change) between two or more samples is the quantity of interest. We concentrate on relative quantification here, since it is deemed much more accurate than absolute quantification [Ong and Mann, 2005]. The most widely used relative quantification techniques rely on the intensities in the MS spectra. This can either be done within a single MS run after samples have been labeled or across runs in a label-free experiment and involves finding intensities that belong to the same peptide in the two samples, a proper way to compute the ratio of all corresponding intensities and normalization. After peptide fold changes are available, they are assembled into protein quantifications. This is usually done for so called protein groups that contain those proteins from the database that share the majority of their peptides [Nesvizhskii and Aebersold, 2005]. The output of such workflows therefore consists of a list of protein groups together with identification statistics and a summarized relative quantification.

When looking at individual peptide fold changes of typical high-throughput mass spectrometry experiments, it becomes clear that in several cases, peptides seem to exhibit a different fold change than other peptides from the same protein (see for instance Figure 7.1). There are several possible explanations for such situations, including:

1. Measurement imprecision: Repeated independent measurements of the same quantity (i.e. peptide fold change) are subject to noise. The variance of the seven independent measurements of the left most peptide in Figure 7.1 for instance are most likely the effect of noise.
2. Ambiguous peptides: The sequence of the left most peptide may not be unique to this protein and its true fold change in the sample should be intermediate between all matching proteins.
3. Wrong identification: An MS² spectrum may erroneously be assigned to a given peptide and the measured fold change therefore belongs to a peptide from a different protein.
4. Wrong quantification: There may be certain properties of peptides that introduce bias into quantification and the normalization of the quantification algorithm may not have corrected for that. For instance, if a peptide of an abundant protein can be ionized easily, saturation effects may lead to underestimated fold changes.
5. Differentially regulated post-translational modifications (PTMs): It is known that post-translational modifications such as phosphorylations are highly regulated and may be

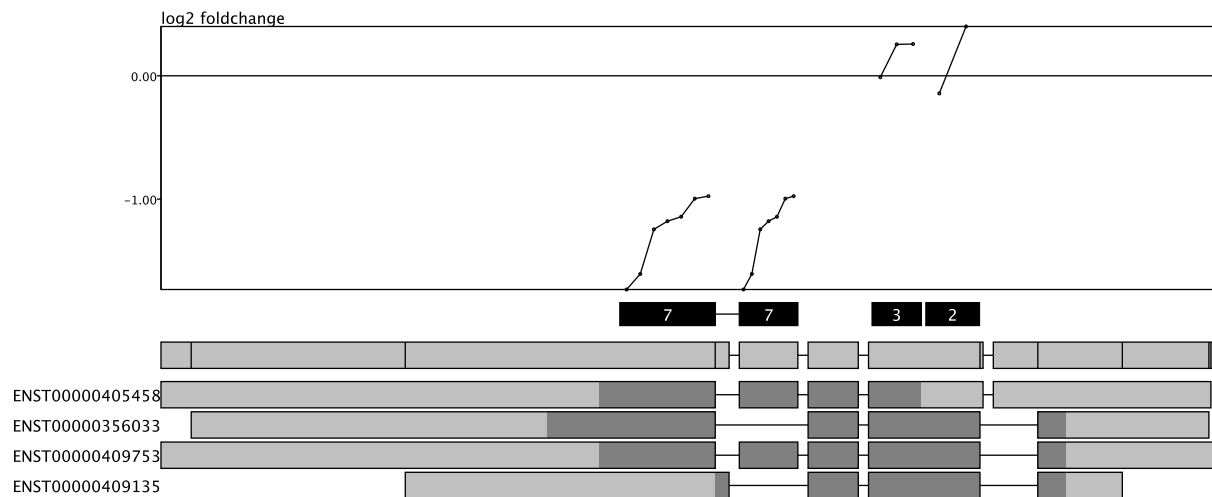


Figure 7.1: Example MS peptide quantifications for a gene with several isoforms. Shown are quantitative mass spectrometry measurements for the gene HN1 in a SILAC experiment as produced by MaxQuant using standard parameters. On the top, \log_2 fold changes of all quantifications for this gene are shown. For each event, a dot is drawn on top of the respective peptide and multiple measurements for the same peptide are shown in increasing order. For the left most peptide, that spans an exon-exon junction, its seven measurements are shown twice (above both exons). On the bottom the gene structure according to Ensembl is shown: The first line corresponds to the gene with all its exons and alternative splice donors and acceptors (black lines) and the remaining lines represent the four transcripts with coding parts in dark gray. For clarity, exons are shown in scale whereas introns are shrunk to a fixed size. All shown peptides uniquely map to these locations.

differential in the conditions under consideration. If only the unmodified version of a peptide has been identified and the modification has been upregulated, the unmodified peptide will have a fold change that is different from the gene fold change.

6. Differential regulation of isoforms: Most eukaryotic genes can give rise to multiple isoforms, either by alternative splicing, alternative transcription start sites or combinations of these. Alternative peptides, i.e. peptides that are not part of all isoforms of a gene are expected to show different fold changes, if respective isoforms are differentially regulated.

Depending on the summarization strategy the protein fold change for the gene in Figure 7.1 would either be around 2-fold down regulated or not regulated (when using the median of all measurements or the median of all peptide medians, respectively). In either case, defining a protein fold change may not be appropriate since the situation is obviously more complex. Thus, a method to detect such situations would be of great benefit and would allow to investigate such situations further.

A first attempt into that direction was made in [Forshed et al., 2011], where the correlation coefficient of intensities of peptide from the same gene across multiple conditions was used as

a distance measure for peptides. Hierarchical clustering was then used to either exclude outlier peptides (that are uncorrelated to all other peptides) or to group genes in order to infer isoforms. However, the correlation coefficient is useless when only two conditions are investigated as in standard SILAC experiments. Also, since it directly compares the XIC of peptides between conditions to compute the correlation coefficient, it cannot make use of more sophisticated ways to compute intensity ratios as for instance implemented in MaxQuant [Cox and Mann, 2008]. Furthermore, excluding peptides that are uncorrelated to all other peptides from the same gene may not always be appropriate, since such a peptide may be the only one specific to an isoform (e.g. if it is located on a cassette exon).

Our goal in this study was to provide a method that is able to detect outlier peptides in standard SILAC experiments. The proposed method was rigorously tested on in-silico simulated data, where it could detect outlier peptides with high performance (as measured by an AUC > 0.8) when the true difference was as small as 1.4 fold. The second goal was to investigate reasons for outlier peptides in experimental data: Given we have identified a set of genes like in Figure 7.1, determine which of the reasons from above play a role in this set.

7.3 Materials and methods

7.3.1 Data processing

Experimental data taken from [Cox and Mann, 2008] has been downloaded from ProteomeCommons Tranche, where EGF stimulated HeLa cells were compared to control cells using SILAC. Data has been analyzed using MaxQuant [Cox and Mann, 2008] version 1.2.0.18 (June 2011) against all proteins downloaded from Ensembl v60 (November 2010). Default parameters have been used: Oxidation (M) and Acetylation (N-term) as variable modifications and Carbamidomethylation(C) as fixed modification, reverse peptides as decoy database, matching between runs in a 2min rt window. For all further analyses, we use all unique peptides from evidence.txt (produced by MaxQuant) that contains quantification events of all identified (and matched) SILAC pairs at a FDR of 1% (according to a decoy database approach). To determine uniquely matching peptides, peptide sequences from evidence.txt have been mapped to the human genome using position information obtained via Ensembl Biomart, and only uniquely matching peptides have been retained. Gene definitions also have been taken from Ensembl, with the modification that overlapping genes have been clustered to gene clusters using single linkage (i.e. a peptide mapped to the genome always belongs to a single gene cluster). We will refer to these gene clusters as genes in the following. In order to perform statistical tests on quantifications, we furthermore discard all peptides if less than 3 independent measurements are available.

7.3.2 Detecting outlier peptides

The goal of our method is to distinguish measurement noise from other reasons that lead to peptide fold changes that are different from other measurements from the same gene. This

is based on an important property of typical mass spectrometry experiments: Many peptides are identified and quantified multiple times because experiments have been done in replicates, because peptides may have been measured in multiple gel slices (which may have been used for a fractionation step before mass spectrometry) or in multiple charge states. Since all these quantifications are technically independent from each other, we can use them to estimate the quantitative precision. The goal then is to determine peptides that are different from other peptides from the same gene and where this difference cannot be explained by a high quantification variance.

The most basic algorithm first computes all peptide and gene fold changes p_i and g_k by taking the mean or median of all corresponding measured fold changes. Then, genes are ranked by their maximal absolute peptide-from-gene deviation

$$d_k = \max\{|g_k - p_i| \mid \text{peptide } i \text{ uniquely belongs to gene } k\}$$

Unfortunately, there are two caveats in such a procedure: First, it is difficult to determine a reasonable cutoff without performing permutation tests and second, it inherently assumes that variance due to noise is equal for all peptides in the dataset. This is certainly not true, since the signal-to-noise ratio depends on the expression level of a gene.

Therefore, we also adapted a classical ANOVA procedure: For each gene, we fit the linear model $F_{ij} = g + p_i + \epsilon_{ij}$ to all \log_2 fold changes of a given gene, where F_{ij} is the j th \log_2 fold change of a repeatedly measured peptide of the gene, g is the gene fold change, p_i is the residual peptide fold change and ϵ_{ij} is the noise in measurement i, j . Residual peptide fold changes that are significantly different from 0 indicate that this peptide behaves differently from other peptides from the same gene. Therefore, genes can be ranked using the p-value from an F test or by $\eta^2 = \frac{SS_p}{SS_g}$ from ANOVA (where SS_p is the within peptide sum-of-squares and SS_g is the within gene sum-of-squares), a classical measure for effect size [Cortina and Nouri, 2000].

The ANOVA model estimates noise levels gene-by-gene and, therefore, deals with different signal-to-noise ratios across genes. Unfortunately, the signal-to-noise ratio could not only depend on expression levels of genes, but also on properties specific to peptides (e.g. ionization efficiency). The ANOVA model however assumes equal variance across peptides. We therefore also adapted the heteroscedastic ANOVA from [Krishnamoorthy et al., 2007], which can deal with different variances.

Thus, we propose five methods to rank genes: Mean distance and Median distance corresponding to ranking by the maximal peptide-from-gene deviation, ANOVA F test p-value and ANOVA η^2 using the classical ANOVA approach and the heteroscedastic ANOVA p-value. For further analyses, we define the outlier peptide of a significant gene as the peptide that has the greatest absolute difference between its \log_2 fold change median and the \log_2 fold change median of the gene. Note that there may be multiple peptides that have fold changes differing from the gene fold change but for simplicity we only used a single peptide per gene.

7.3.3 In-silico data generation

For the experimental data, no standard of truth is known, i.e. there is no knowledge about differentially regulated isoforms between stimulated and control HeLa cells. Therefore, we

simulated mass spectrometry data in-silico which allowed us to provide a controlled environment for a rigorous evaluation of our method. Instead of attempting to simulate the physical events in a mass spectrometer, we chose to directly generate peptide quantifications, which is the type of measurement our method works on. We use experimental data (see Data processing) to estimate model parameters for the simulation and, therefore, the simulated data has the same properties as experimental data. As a consequence, evaluation results for simulated data should also apply to experimental data.

We proceed as follows: We consider each Ensembl gene with at least two isoforms. First we draw the number of measured peptides for a gene and distribute these peptides across all isoforms. We discard these peptides and repeat this step if there is no specific peptide, i.e. a peptide that is not present in at least one isoform. Then we set the isoform \log_2 fold changes f_0 and f_1 depending on whether we want to generate a positive or a negative example. For positive examples, we set $f_0 = 0$ and $f_1 = f > 0$, for negative ones we set $f_0 = f_1 = 0$. Then, for each peptide p , we draw the number of measurements n and the variance σ^2 based on the empirical distributions obtained from the experimental data. $n \log_2$ fold changes for p are drawn according to $N(\mu, \sigma^2)$, where $\mu = \frac{I_0 \cdot f_0 + I_1 \cdot f_1}{I_0 + I_1}$, where I_i is an indicator variable for peptide p to be contained in isoform i . Therefore, μ is either f_0, f_1 or their arithmetic mean, depending on the location of the peptide (unique to isoform 0, unique to isoform 1 or on a shared exon).

Thus, we can generate N positive examples for a defined fold change value f and N negative examples. Each positive example represents a gene that is not regulated transcriptionally but have isoform proportions differing by a factor of f (e.g. by differential regulation of alternative splicing). Negative examples represent genes that are not regulated at all. Our methods are able to compute a score for each gene and therefore, we can use these N positive and N negative examples to evaluate their accuracy using ROC curves. Since the number of peptides and measurements per peptide and the quantification noise is drawn according to distributions from experimental data, and only the isoform fold change difference is set by hand, results from this in-silico evaluation can be expected to apply to experimental data also.

7.4 Results and Discussion

In a typical high-throughput quantitative mass spectrometry experiment, hundreds of thousands of precursor ion measurements can be used for peptide quantification. Usually, a single peptide is detected and quantified multiple times either due to biological or technical replicates or to repeated measurements within a single replicate in different charge states, different gel slices etc. (see Figure 7.2 for the SILAC data from [Cox and Mann, 2008]). As introduced above, there are several reasons why peptides may have been measured with differing fold changes even if they are products of tryptic digestion of the same protein.

If we assume a complete protein database, excluding ambiguous peptides is straight-forward (see Methods). And even if the database is not complete, the likelihood that an outlier peptide also occurs in another unknown but expressed protein is negligible. We want to emphasize that we only treat peptides matching to multiple genomic locations as ambiguous. For instance, peptides

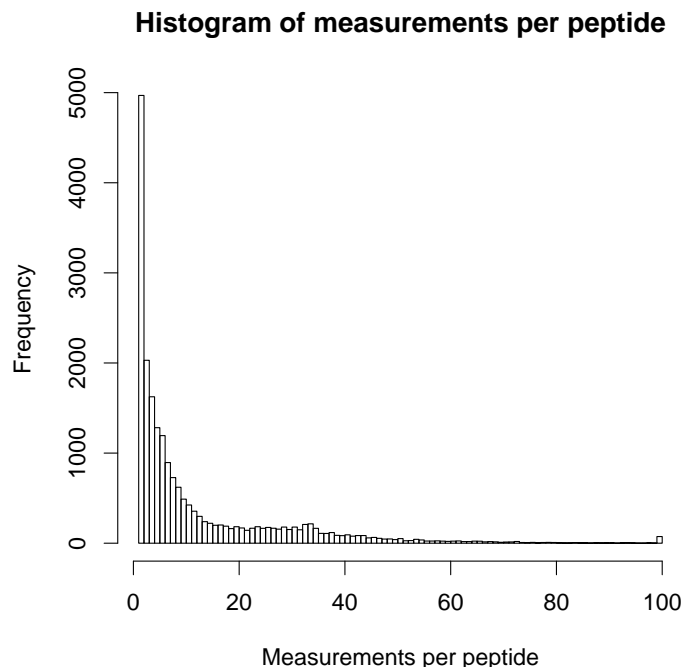


Figure 7.2: Number of measurements per peptide. The figure shows a histogram for uniquely matching peptides for a dataset taken from [Cox and Mann, 2008] and processed using MaxQuant with default parameters (see Materials and Methods for details). Overall, 265000 peptide measurements out of 344000 are shown in the histogram. For clarity, all counts > 100 have been set to 100.

coming from constitutive exons of an alternatively spliced gene (which occur multiple times in our protein database) are still unique by our definition.

Thus, the main focus of our algorithm is to distinguish noise from other reasons for outlier peptides. In order to rigorously test the accuracy of the proposed methods, we applied them on in-silico generated data for which we know the true situation. This allowed us to circumvent the problem of missing gold standards. We furthermore applied our algorithm to real data in order to delineate which reasons other than noise can lead to outlier peptides.

7.4.1 Test on in-silico generated data

We simulated peptide quantification datasets for several true fold change differences f (see Methods) and for all methods proposed and evaluated them using ROC curves and the AUROC (see Figure 7.3). According to the AUROC scores in Figure 7.3a, all methods seem to behave very similar across the whole range of true fold changes. When looking at individual ROC curves however, we note that their performance at different score cutoffs is quite different: Gene-wise variance estimation seems to perform much better at high specificity score cutoffs, whereas

experiment-wide variance estimation has higher sensitivity at lower cutoffs (see also Figures 7.3c and 7.3d). For all further analyses, high specificity is important, and we will therefore use an ANOVA procedure in the following, which also allows us to compute a statistically sound cutoff. As can be seen in Figure 7.3d, for a p-value cutoff of 0.01, the heteroscedastic ANOVA performs superior independent of the true fold change difference and thus we propose this method as the method of choice.

We note that the fold changes reported in Figure 7.3a already account for the fact that the fold change difference between a specific peptide and a constitutive peptide is expected to be smaller than the isoform fold change, so the peptide fold change that is enough to detect significant differences between peptides is actually even lower than 0.5 on \log_2 scale.

We acknowledge that testing a method on in-silico generated data can lead to overoptimistic conclusions: If the model that generates the data is oversimplified, an oversimplified method's performance would be overestimated. One main goal of our model is to test for influence of in-gene heteroscedasticity of quantifications. Since our model generates unequal variances in a realistic way by using the variance distribution obtained by real data, our generated data is affected by heteroscedasticity to the same extent as experimental data. There are two possible outcomes of our evaluation: If the standard ANOVA would perform equal or even superior to the heteroscedastic ANOVA, then heteroscedasticity is not an important factor and respecting it is not justified due to a greater power of the standard ANOVA in a homoscedastic situation.

However, we observe that a test that respects possible unequal variances performs better than tests that assumes homoscedasticity and thus we can conclude that heteroscedasticity indeed plays an important role in such kind of data and that it is beneficial to use our heteroscedastic ANOVA for real data. Furthermore, we observe that we are able to detect differentially regulated isoforms reasonably well (as judged by an $AUC > 0.8$) if their fold change is as small as ~ 1.4 fold (i.e. the \log_2 fold change is 0.5), if we observe at least one specific peptide (i.e. a peptide that is not part of one of the differential isoforms).

7.4.2 Outlier peptides in real data

We applied our method based on the heteroscedastic ANOVA on experimental data taken from [Cox and Mann, 2008]. As can be seen in Figure 7.4, there are several genes that have significantly different peptides and their fold change distribution is as expected by our in-silico data analysis: The majority of genes shows a $\geq 0.3 \log_2$ fold change which matches the performance measured by our ROC analysis (see Figure 7.3).

We next made an attempt to reveal why these peptides show a fold change that is different from the gene fold change. For the following analyses, we used an (uncorrected) p-value cutoff of 1% in order to get a reasonable large set of peptides that is still enriched with real differential peptides. By setting this cutoff, from the 3314 genes, we extracted 257 peptides (we will refer to them as outlier peptides). We extracted a background set of 1850 peptides from genes with a p-value of > 0.5 .

First, we checked whether there is an indication of misidentifications within our outlier peptides. To this end, we checked whether there was a second best candidate in the list of identifications for the corresponding MS^2 spectrum and extracted its score if another candidate peptide was found.

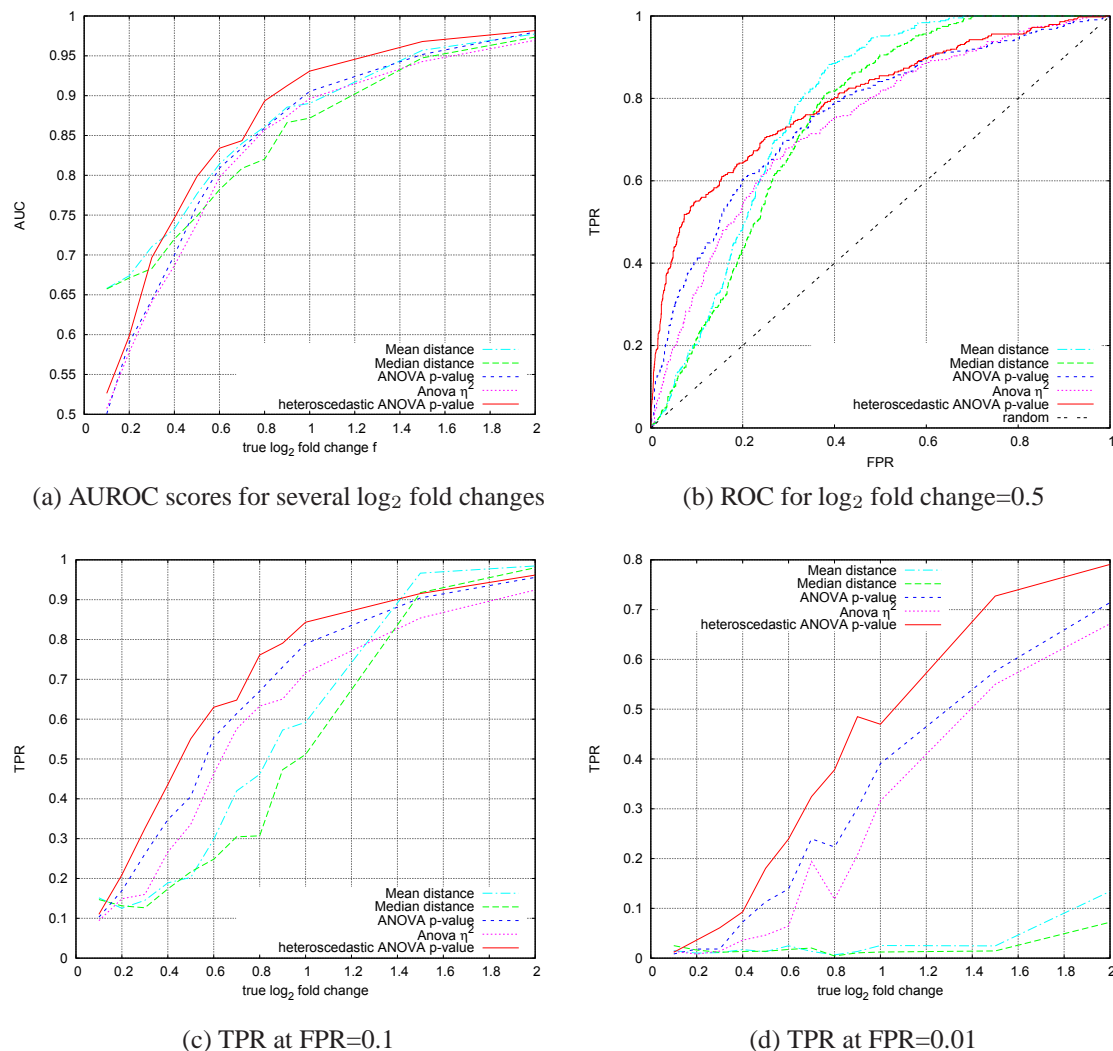


Figure 7.3: Evaluation on in-silico generated data. For several true fold changes f ranging from 0.1 to 2, 1000 positive and 1000 negative examples have been generated (see Methods). Genes were then ranked according to the five proposed methods and ROC curves together with their area under the ROC curve (AUROC) were computed. Shown in 7.3a is the AUROC value for all computed ROC curves and all proposed methods. In 7.3b the ROC curve for $f = 0.5$ is shown. Figures 7.3c and 7.3d show the true positive rates for a fixed false positive rate of 0.1 and 0.01, respectively, and emphasize the superiority of the ANOVA procedures in comparison to the simple methods at high specificity cutoffs.

This revealed that the outlier peptides have statistically significantly more additional candidate peptides than expected by our background peptides ($p = 0.0073$, Fisher's exact test on the number of peptides that have only single candidate spectra; $p = 0.0018$, Kolmogorov-Smirnov

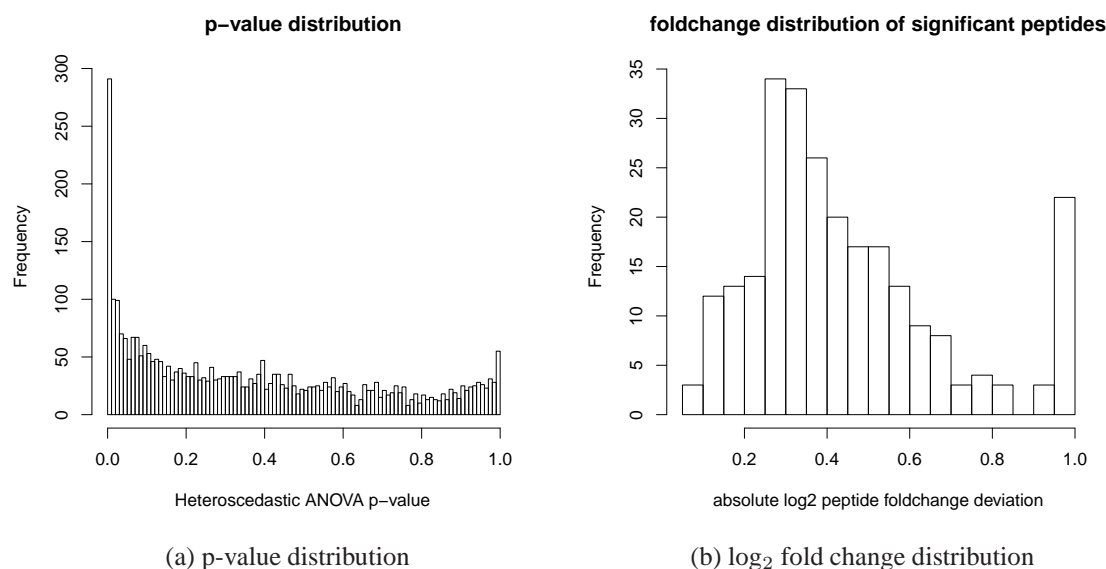


Figure 7.4: Heteroscedastic ANOVA applied to the experimental data. Shown is the distribution of all p-values in 7.4a and in 7.4b the \log_2 distribution of all significant peptides. For clarity, in 7.4b, all values > 1 have been set to 1.

test on the fraction of quantification events for a peptide having additional candidates; see also Figure 7.5a).

This means that even if all these peptides have been independently identified multiple times, there is evidence that in several cases, all these independent quantifications erroneously are assigned to the same peptide. A reason for that could be that some peptides in the proteome are very similar to each other, either directly in their sequence or with respect to additional unknown properties that lead to a similar fragmentation pattern. This is also directly reflected in the scores of the peptide candidates: An Andromeda score is $-\log_{10}(p)$ of a p-value p testing the Null hypothesis that a peptide does not belong to a given MS^2 spectrum. There are several cases where multiple candidates have a score > 10 and if we assume that only a single peptide species has been chosen for fragmentation, all but one of these scores are overestimated. It is a-priori not clear, if the top candidate necessarily is always the correct one.

We also noted that sometimes there were extreme outliers within the independent quantification events of a peptide as judged by an interquartile range (IQR) distance of > 1.5 . When we performed similar tests on these IQR outliers compared to all quantifications within the IQR, we also observe statistically significant more additional candidates than expected by background ($p < 10^{-26}$, Fisher's exact test on the number of quantification events that have additional candidates; $p < 10^{-11}$, Kolmogorov-Smirnov test on the ratio of second score to the best score, see also Figure 7.5b).

Then, we tested whether there is bias with respect to several physico-chemical properties. These properties have been taken from [Mallick et al., 2007], where they have been used to predict

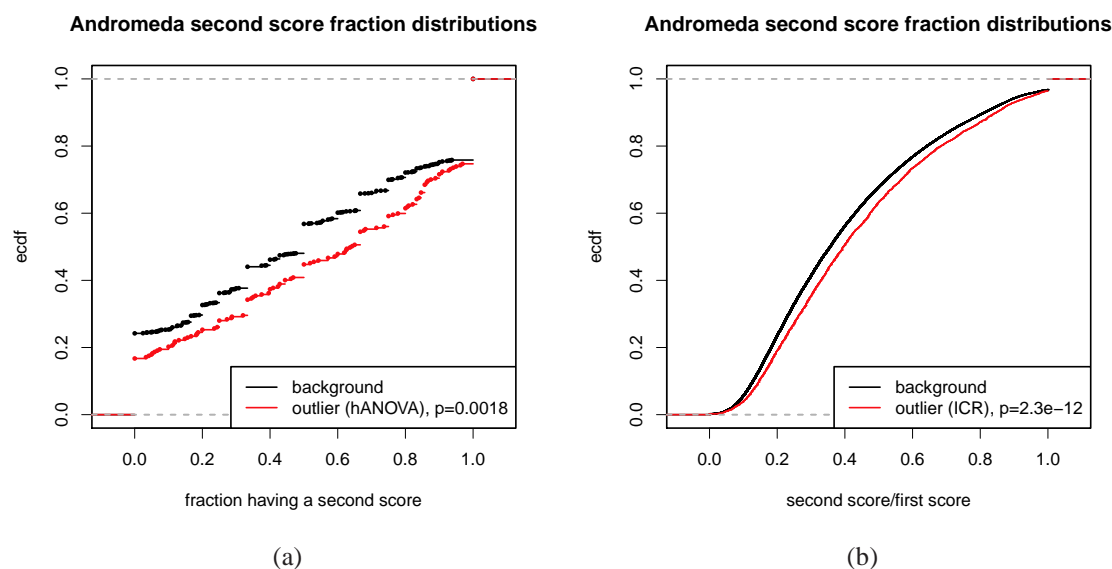


Figure 7.5: Evidence for misidentifications in outlier peptides. Figure 7.5a shows the distributions of the fraction of spectra that have multiple candidates for outlier and background peptides. E.g. a peptide where 5 out of 10 measurements have additional candidate peptides would be counted with the value 0.5. For instance, about 55% of the background peptides have a value of ≤ 0.5 in comparison to about 45% for outlier peptides, which is a statistically significant difference according to a Kolmogorov-Smirnov test. In Figure 7.5b the distributions of the fraction of second to best candidate score for outlier measurements and background measurements. Here, outliers are not defined by the ANOVA but by an interquartile range distance of > 1.5 . E.g. if there is a peptide with 10 measurements, we can calculate the interquartile range r as the difference between (sorted) measurements 2 and 9. If the difference between measurements 1 and 2 or 9 and 10 are $\geq 1.5r$ then they are deemed outliers, respectively. Such outliers have better second candidates than all other measurements.

proteotypic peptides. Each of these properties allows to compute a score for a given peptide sequence. For each property, we computed scores for all outlier peptides and all background peptides and compared the score distributions by a Wilcoxon-Mann-Whitney test. The p-value distribution of these tests shown in Figure 7.6 clearly shows that most of these physico-chemical properties are significantly different between outlier peptides and background peptides. This means that the normalization used by MaxQuant is not able to correct for bias introduced by these properties. It should however be noted, that several of these properties are not independent, for instance there are several properties that try to measure hydrophobicity. One interesting example (which is directly related to hydrophobicity) is the striking difference in retention times ($p < 10^{-5}$, Wilcoxon test). This analysis shows that there are more outlier peptides with short retention time than expected, which is probably only due to technical bias that should be removed by further normalization.

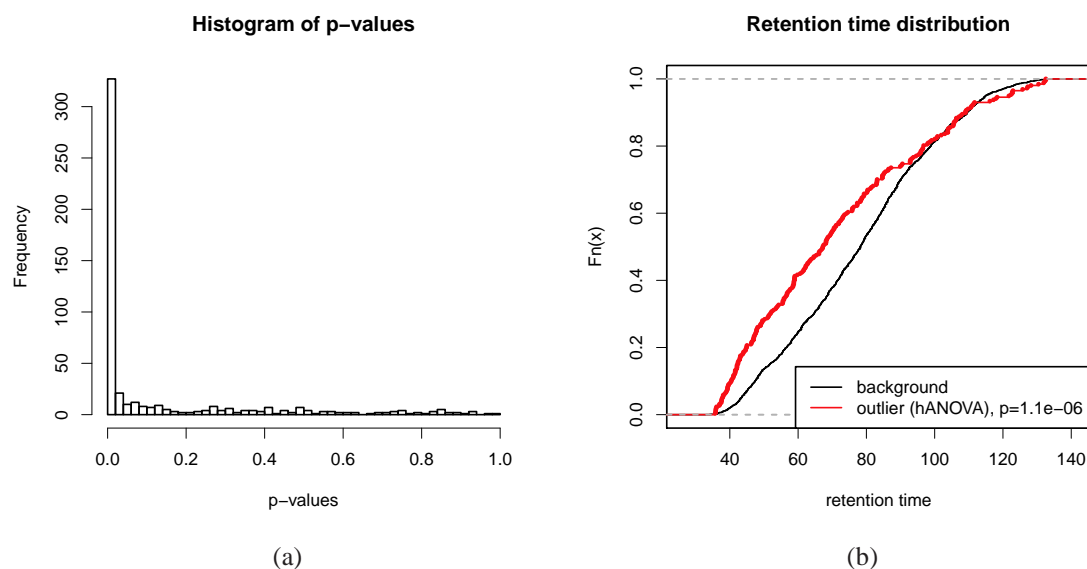


Figure 7.6: Evidence for misquantifications in outlier peptides. Shown is an histogram of the p-values of all physico-chemical properties tested in 7.6a and the cumulative distributions of retention times for outlier peptides and background peptides in 7.6b. See text for details.

Another possible explanation for misquantification is saturation, which means that for extremely abundant peptides, reported intensities may be underestimated. When, for instance, two peptides from the same protein have differing ionization efficiencies, computed fold changes may be different due to this saturation effect. And indeed, outlier peptides have higher intensities than expected by background ($p < 10^{-13}$, Kolmogorov-Smirnov test), which indicates that saturation is another effect that should be removed by proper normalization.

We also made an attempt to test for differential post-translational modifications. Allowing phosphorylation as a variable modification during peptide identification in Andromeda yielded only very few results and the correctness of these identifications should be doubted (data not shown). This however was expected since in the dataset we used, phosphopeptides have not been enriched experimentally. However, the absence of reliably identifiable phosphopeptides does not prove their absence in the sample: If without enrichment the phosphopeptides abundance in the mass spectrometer is lower than the unmodified peptide, it will not be selected for fragmentation and MS^2 . Thus, we downloaded known phosphopeptides from a publicly available database [Bodenmiller et al., 2008] and tested whether there is an overlap of these peptides with our outlier peptides. Even if there was only a small number of phosphopeptides detected in our experiment and it is not clear if they are also phosphorylated here, there was a weak but statistically significant overlap ($p = 0.034$, Fisher's exact test). This means that differential PTMs indeed seem to be present in our dataset and that they can be detected using our method.

Finally, we tried to find evidence for differentially regulated isoforms in our dataset. To this end, we classified each peptide location as alternative or constitutive location. Due to the sparseness

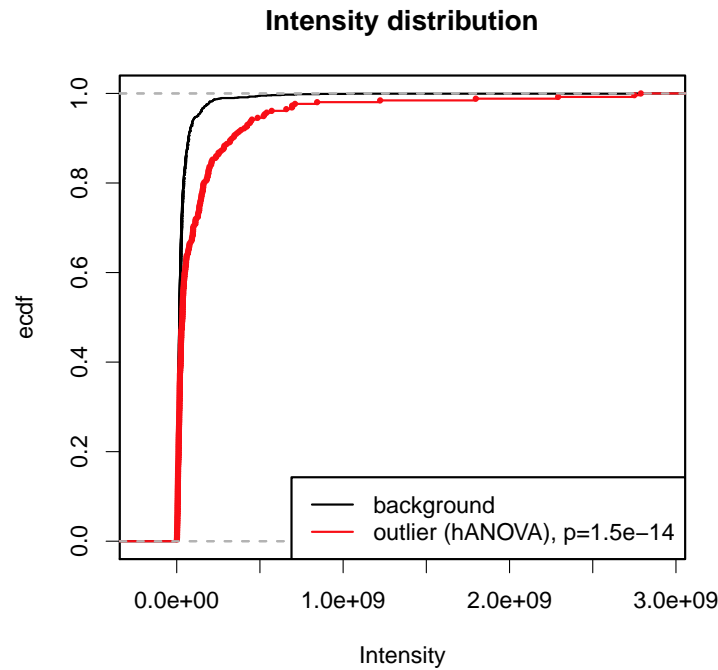


Figure 7.7: Evidence for saturation in our dataset. The distributions for summed intensities for all detected peaks is shown for outlier peptides and for background peptides.

of the identified peptides, it is impossible to infer which isoforms are expressed in our data and therefore we cannot restrict the transcripts to expressed transcripts. Thus, we classify based on the full Ensembl annotations: A constitutive peptide is contained in all Ensembl transcripts of the corresponding gene and a peptide is characterized alternative if there is at least one transcript that does not contain the peptide. Surprisingly, we found a small but statistically significant enrichment of outlier peptides among constitutive locations ($p = 0.0086$, Fisher's exact test), which supposedly suggests that background and not outlier peptides are parts of differentially regulated isoforms. However, we noted that the exon length (defined as the number of nucleotides in a gene, that is part of at least one Ensembl exon) is significantly larger for our outlier peptides than for our background peptides ($p < 10^{-16}$, Kolmogorov-Smirnov test). Thus, we removed this bias from the analysis by sampling peptides from our background set according to the exon length distribution of our outlier peptides. When we apply the same test as without sampling, outlier peptides are now enriched among alternative location. This enrichment is however not statistically significant ($p = 0.3$, Fisher's exact test), which is either a consequence of the small numbers or indeed true: Probably in our dataset, differential regulation of isoforms is not as widespread as all the other effects, in which case a statistic over the whole set of outliers is not expected to yield significant results.

7.4.3 Discussion

Our results demonstrate that all effects introduced before may be present in the set of outlier peptides. Our main future goal is to be able to distinguish between these effects: Errors (misidentifications and misquantifications) could be reduced by improving both identification algorithms and normalization methods. Detecting outlier peptides can help to do that: For instance, if an identification algorithm has to choose between multiple candidate peptides for a spectrum, it could use the outlier score as an additional criterion to do so. It also seems as if the normalization in MaxQuant, that accounts for intensity, labeled amino acids and different protein load [Cox and Mann, 2008], is not able to remove all bias from the data.

Post transcriptional modifications (PTMs) have received increasing interest in recent years [Huttlin et al., 2010; Olsen et al., 2010]. Usually, specific steps during sample preparation are made to enrich modified peptides such that they can readily be detected and identified. We have shown that even without these enrichment steps, differential PTMs are in principle detectable in a standard MS experiment, even if peaks corresponding to the modified version of a peptide are not selected for fragmentation. Our outlier peptide score can be used to generate hypotheses for finding differential PTMs.

Alternative isoforms, which are consequences of alternative transcription start sites, alternative splicing or alternative end-of-transcription sites (or combinations of these), are widespread in higher organisms [Pan et al., 2008] and it is known that they are highly regulated in development [Chawla et al., 2009; Cooper, 2005; Lynch, 2004], between different tissues [Wang et al., 2008] and in diseases [Cooper et al., 2009; Grosso et al., 2008]. Experimental techniques to detect differentially regulated isoforms usually only consider isoforms on the mRNA level [Pan et al., 2008; Wang et al., 2008]. However, it is known that not all produced transcripts give rise to an equal number of proteins, so the ultimate test for differential regulation of isoforms must be performed on the protein level.

Finding differentially regulated isoforms is thus probably the most interesting application of our method, even if we were not able to reliably find cases in the dataset we used. To our knowledge, there is no established method available that has the ability to detect differential isoforms on proteome level in a high-throughput manner. Once other effects can be excluded for an outlier peptide, quantitative mass spectrometry could serve this purpose: The only explanation that remains for outlier peptides then is indeed differential regulation of isoforms. Furthermore, we can expect that in the future, the number of identified peptides will increase due to technical progress and due to improved computational methods [Cox and Mann, 2008]. Even if there certainly are peptides that are not detectable in mass spectrometers, the number of peptides that can nowadays be identified is orders of magnitude lower than what is actually quantifiable in modern mass spectrometers [Michalski et al., 2011]. Due to technical and computational advances, we expect that in the near future, the protein coverage by peptides will provide a more complete picture. This will also help to distinguish differentially regulated isoforms from the other effects, since then, regularly more than one quantified peptide will be specific for isoforms.

7.5 Conclusion

In modern quantitative high throughput mass spectrometry data, the final analysis step is to compute protein fold changes for all identified proteins. In most cases, this seems to be straight-forward as long as a robust statistic is used to compute the protein fold change from all individual quantifications. However, when having a closer look at individual peptide quantifications, it becomes evident that protein fold changes are not always a proper way to summarize measurements. In many cases there are peptides that are significantly different from other peptides of the same gene. This could be because the peptides stem from alternative isoforms of a gene and because respective isoforms are differentially regulated in the conditions under consideration. We propose a method that is able to detect such differential regulation of isoforms.

However, we found several effects that could confound this explanation of the quantifications in real data: misidentifications, misquantifications and post-translational modifications. Unfortunately, it is a-priori not clear which of these effects plays a role for a gene. Thus, in order to reliably detect differentially regulated isoforms and distinguish it from these effects, additional data is necessary. If for instance RNA-seq data is available for the same cells used for mass spectrometry, it could provide additional evidence for differentially regulated isoforms by sequencing reads that support these isoforms either qualitatively or even quantitatively.

This study also revealed that the normalization currently used is not sufficient to remove all technical bias. For instance, we have shown that the retention time (either directly or something that is correlated with it) affects quantification and further normalization is necessary to remove this bias. Our method is able to provide peptides that are probably affected by such a bias which can help in the development of further normalization steps.

In a modern mass spectrometer, only a limited number of all the peptides detectable in MS spectra is selected for fragmentation and MS² [Michalski et al., 2011]. In order to find differentially regulated isoforms, it would be beneficial to increase the number of identified peptides: Usually there is more than one peptide specific to a single isoform. If multiple specific peptides are detected and measured, all other effects as described above become less probable. Due to the increasing throughput and decreasing scan times, we expect that such kind of data will be available soon and our method could then even better be used to systematically search for differentially regulated isoforms.

Chapter 8

FERN - Stochastic Simulation and Evaluation of Reaction Networks

Motivation: *The previous chapters were either about raw data analysis (chapters 3-5) or the interpretation of systems biology data with respect to specific research questions, namely context-dependence of microRNA-mediated regulation in chapter 6 and differential splicing in chapter 7. In this chapter, I complete the workflow described in the introduction (see section 1.1.1) by describing a software package for stochastic simulation of biological networks. In my diploma thesis, I developed the Petri Net Modelling application (PNMA), a comprehensive modelling platform for biological networks, which provides a highly flexible system for simulations of such networks and originally provided methods to simulate networks using Fuzzy logic systems that are powerful when experimental data is sparse and noisy. When detailed experimental data is available, a system can be modeled in more detail and more detailed simulations are possible. A widely used, highly detailed model is based on stochastic simulation, where reactions of individual molecules are considered. Thus, I developed a Java library for stochastic simulation, which I integrated into PNMA, but which is also available for the widely used software packages Cytoscape and Celldesigner.*

Publication: *FERN has been published in BMC Bioinformatics [Erhard et al., 2008] and in this extended form as a book chapter by Springer [Erhard et al., 2010]. Here, I adapted the layout and made minor corrections to the text of the book chapter.*

My contribution: *I developed and implemented the software, carried out evaluations, drafted the paper and wrote the extensions for the book chapter text.*

Contribution of co-authors: *Caroline Friedel helped to draft the original manuscript. Ralf Zimmer supervised the work and helped to revise the manuscript.*

8.1 Abstract

Stochastic simulation can be used to analyze the development of biological systems over time and the stochastic nature of these processes. Most available programs for stochastic simulation, however, are limited in that they either a) do not provide the most efficient simulation algorithms and are difficult to extend, b) cannot be easily integrated into other applications or c) do not allow to monitor and intervene during the simulation process in an easy and intuitive way. Thus, in order to use stochastic simulation in innovative high-level modeling and analysis approaches more flexible tools are necessary. FERN (Framework for Evaluation of Reaction Networks) is a Java framework for the efficient simulation of chemical reaction networks. It is subdivided into three layers for network representation, stochastic simulation and visualization of the simulation results each of which can be easily extended. It provides efficient and accurate state-of-the-art stochastic simulation algorithms for well-mixed chemical systems and a powerful observer system, which makes it possible to track and control the simulation progress on every level. To illustrate how FERN can be easily integrated into other systems biology applications, plugins to Cytoscape and CellDesigner are included. These plugins make it possible to run simulations and to observe the simulation progress in a reaction network in real-time from within the Cytoscape or CellDesigner environment. FERN addresses shortcomings of currently available stochastic simulation programs in several ways. First, it provides a broad range of efficient and accurate algorithms both for exact and approximate stochastic simulation and a simple interface for extending to new algorithms. FERN's implementations are considerably faster than the C implementations of gillespie2 or the Java implementations of ISBJava. Second, it can be used in a straightforward way both as a stand-alone program and within new systems biology applications. Finally, complex scenarios requiring intervention during the simulation progress can be modelled easily with FERN.

8.2 Background

Traditionally, wet-lab experiments were focused on describing the function of individual genes or proteins. With the advent of high-throughput technologies, system-level approaches have become common, which make it possible to identify the interactions between the individual elements of the cell. Here, mathematical models are crucial in understanding these biological systems. In particular, the simulation of the dynamics of these can visualize and predict quantitative aspects of the system such as gene expression in regulatory networks or signal amplification in signal transduction networks [Szallasi et al., 2006].

The most common approach to modeling dynamics is via ordinary differential equations (ODEs) which describe deterministically how the system evolves with time (see e.g. [Clodong et al., 2007; Shimon et al., 2007; Calzone et al., 2007]). Since the simulation of ODEs is deterministic, successive simulations starting from the same initial conditions lead to the same results. Many aspects of biological systems are not deterministic, which can lead to quite different outcomes for the same initial conditions. In addition, when small numbers of molecules are involved, concentrations of the involved molecules cannot be considered to be continuous, which is one of

the fundamental assumptions for ODE models. To address these problems, stochastic simulation algorithms (SSAs) have been developed.

SSAs operate on reaction networks, graph structures containing the molecular species and reactions as vertices and their wiring as edges. These reaction networks are introduced in the following in terms of the more general framework of Petri nets.

8.2.1 Petri nets

Petri nets [Murata, 1989] are a well-established framework for modeling concurrent systems. They do not only provide a variety of analysis tools for biological networks [Reddy et al., 1993] but are also able to simulate their dynamic behaviour in an intuitive way. Furthermore, they provide a straight-forward graphical representation, which makes it possible to model biological networks interactively and to observe or even control simulations. The reaction networks proposed by Gillespie in his original work [Gillespie, 1976] are a special case of Petri nets and are also often called Stochastic Petri nets (SPN). As Gillespie did not use the Petri net nomenclature, his original notations are given in parentheses in the following definition:

Definition 1. A Petri net (reaction network) is a 5-tuple $PN = (P, T, F, W, M_0)$:

- P is a finite set of places (molecular species)
- T is a finite set of transitions (reactions)
- $P \cap T = \emptyset$
- $F \subseteq (P \times T) \cup (T \times P)$ is a set of arcs
- $W : (P \times T) \cup (T \times P) \rightarrow \{0, 1, 2, 3, \dots\}$ is a weight function (stoichiometry)
- $M_0 : P \rightarrow \{0, 1, 2, 3, \dots\}$ is the initial marking of tokens (molecule population)

Thus, a Petri net is a bipartite graph composed of two types of nodes: places (P) and transitions (T), connected by directed edges called arcs (F). Arcs from a place p to a transition t are called *input arcs* of t . p is then called *input place* of t . Arcs from transition t to place p are *output arcs* of t and p is an *output place* of t . Places are marked with a number of tokens according to the marking M and each arc f has an associated positive integer weight $W(f)$. Note that $W(f) = 0 \Leftrightarrow f \notin F$.

A transition t is enabled, if $M(p) \geq W(p, t)$ for each input arc, i.e. each input place is marked with at least as many tokens as the input arc weight requires. The *firing* of an enabled transition t updates markings of connected places: The marking of each input place p_i is set to $M(p_i) - W(p_i, t)$ and each output place p_o is updated with the value $M(p_o) + W(t, p_o)$. This means that a firing transition removes tokens from each input place and adds tokens to each output place.

The graphical representation of a Petri net is shown in Figure 8.1. Places are depicted as circles, transitions as rectangles. Furthermore, the firing of the transition and the update of tokens is shown.

From this intuitive definition, a simple mathematical description of the state and of a firing transition can be derived: If the ordering of $P = \{p_1, p_2, \dots, p_n\}$ is clear, the *current state* of a Petri net can be described by a vector $S = (s_1, \dots, s_n)$ where $s_i = M(p_i)$. Furthermore, each transition t_i can be described by a *state change vector* v_i with components

$$v_{ij} = W(t_i, p_j) - W(p_j, t_i) \quad (8.1)$$

Hence, $S + v_i$ is the state after firing transition t_i . In Figure 8.1, the places are $P = (H_2, O_2, \text{Water})$ and the initial state vector is $S = (3, 1, 1)$. Firing of the transition with state change vector $v = (-2, -1, 2)$ results in the state $S + v = (1, 0, 3)$.

A *simulation* of a Petri net is a sequence of firing N arbitrary transitions. Each firing updates the state vector $S^k \xrightarrow{t_j} S^{k+1}$ according to the state change vector of the firing transition t_j . The resulting sequence of state vectors S^0, \dots, S^N is called *trajectory*. The order of firing transitions is determined by a *firing rule*. A generic simulation algorithm is given in Listing 8.1.

```
while (there is an enabled transition) {
  t = choose an enabled transition according to firing rule
  fire t
}
```

Listing 8.1: Generic Petri net simulation

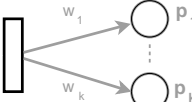
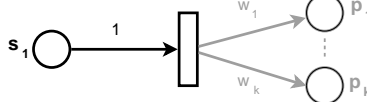
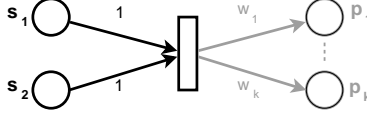
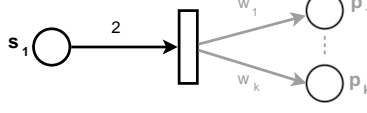
As indicated above, a single token on a place p represents a single molecule of the molecular species p . A firing transition represents the occurrence of a chemical reaction. As a consequence, a trajectory fully describes the temporal dynamics of a molecular model. Thus, to simulate a biological system according to stochastic chemical kinetics, an appropriate firing rule must be derived.

8.2.2 Stochastic chemical kinetics

For our purpose, we assume a well-mixed system with a homogeneous distribution of molecules in a fixed volume at a constant temperature. In this case, many nonreactive molecular collisions will occur until eventually some colliding molecules react. Since we assume a well-mixed system, it is not necessary to track positions and velocities of individual molecules and only the amount of each molecular species and their reactive collisions in the system must be considered. The amount of each molecule is stored as the number of tokens on the respective place in the model and a reactive collision corresponds to a firing transition.

Each state S_l of a trajectory is thus considered as random variable and assigned a time t_l . Hence, $P(S_l, t_l)$ is the probability of having state S_l at time t_l . For each reaction r_i there exists a *propensity function* a_i such that $a_i(S)dt$ is the probability that there occurs a reactive molecular collision of input molecules of reaction r_i in the next infinitesimal time interval $[t, t + dt)$ [Gillespie, 1992]. Then the probability of having state S at time t given the initial state S_0 at t_0 is given by the chemical master equation (CME):

Table 8.1: Mass action propensity functions for basic reactions. c is the respective reaction rate constant. Note, that the propensity function only depends on its input places and not on its output places.

Reaction	Propensity function
	c
	$c \times s_1$
	$c \times s_1 \times s_2$
	$c \times \frac{s_1 \times (s_1 - 1)}{2}$

$$\frac{\partial}{\partial t} P(S, t | S_0, t_0) = \sum_{j=1}^{|T|} a_j(S - v_j) P(S - v_j, t | S_0, t_0) - \sum_{j=1}^{|T|} a_j(S) P(S, t | S_0, t_0) \quad (8.2)$$

This equation describes the time evolution of the probability of having state S at time t for fixed initial conditions. In order to get to S , either one of the $|T|$ reactions has to fire in the last infinitesimal time interval or none. If reaction j fires, the system must be in state $S - v_j$ before the firing (this is the first sum in the CME). If none of the $|T|$ reactions fires, the system must already be in state S (second sum). For a more elaborate description of the CME, see [Gillespie, 1992].

The CME fully describes the stochastic dynamics of a system if the propensity functions for each transition in the model is given. The most prominent form of propensity function comes from mass action kinetics [Gillespie, 1976]. See Table 8.1.

Note that any higher order reaction can be built by composing these zero-, first- and second-order functions. For example, the mass reaction propensity function a of the reaction in Figure 8.1 is $a(S) = c \times \frac{S(H_2) \times (S(H_2) - 1)}{2} \times S(O_2)$.

From the definition of the propensity function, the *next-reaction density function* can be derived, which gives the probability that a specific reaction is the next to occur in an infinitesimal time interval τ :

$$p_i(\tau, j | S, t) = a_j(S) \exp(-a(S)\tau) \quad (8.3)$$

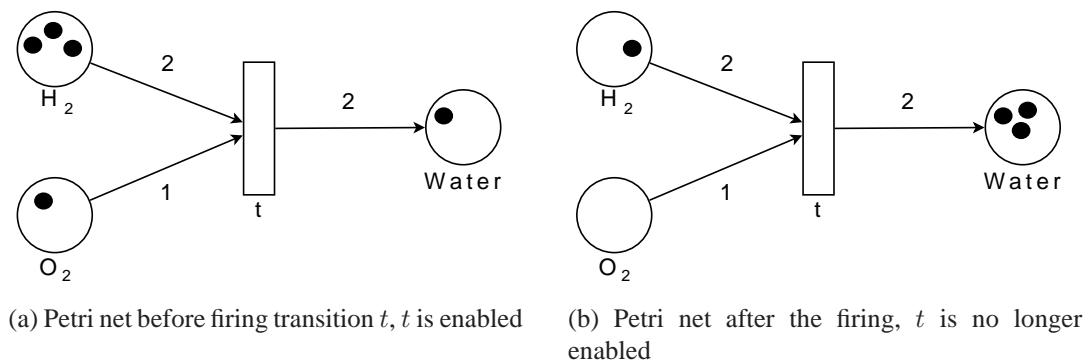


Figure 8.1: A Petri net and the firing of a transition. The Petri net represents the chemical formula $2H_2 + O_2 \rightarrow 2H_2O$. Tokens are shown as filled circles and arc markings as numbers.

where S is the current state vector, τ the time interval and $a(S) = \sum_{j=1}^{|T|} a_j(S)$ the sum of all propensity functions in the system.

A stochastic firing rule which fires transitions according to this density function generates trajectories that have the same probability distribution as the chemical master equation [Gibson and Bruck, 2000]. Usually the CME cannot be solved analytically or numerically, but by sampling a large enough number of SSA trajectories, it can be approximated to an arbitrary precision.

The previous assumptions of constant volume and temperature can be relaxed in this framework. To incorporate changing volume or temperature, the propensity functions are not only dependent on the current state S but also on the current temperature and volume (see [Gibson and Bruck, 2000] for details).

If many reactions can fire without changing propensity functions significantly, the Langevin method [Gillespie, 2001] can be used to describe the stochastic process in a continuous manner in contrast to the discrete form of SSAs. If infinitely large molecular populations are assumed, the Langevin method can in turn be approximated by reaction rate equations, which are a kind of ODEs. Since a trajectory from an SSA run is a sample of the CME, the average of a large sample of stochastic trajectories resembles the solution of the respective ODE.

8.2.3 Stochastic simulation methods

The firing rule introduced in the previous section can be implemented in various ways. Exact methods generate random pairs (τ, j) according to equation (8.3), adjust the state vector by the respective state change vector v_j and advance the time t to $t + \tau$. Then the propensity functions are recalculated to allow generating the next random pair (see Figure 8.2).

First reaction method

The most basic method for drawing random pairs (τ, j) according to equation (8.3) is to generate tentative reaction times τ_i for each reaction r_i according to an exponential distribution with parameter $a_i(S)$ [Gillespie, 1976]. This can be done by using the inversion method, i.e.

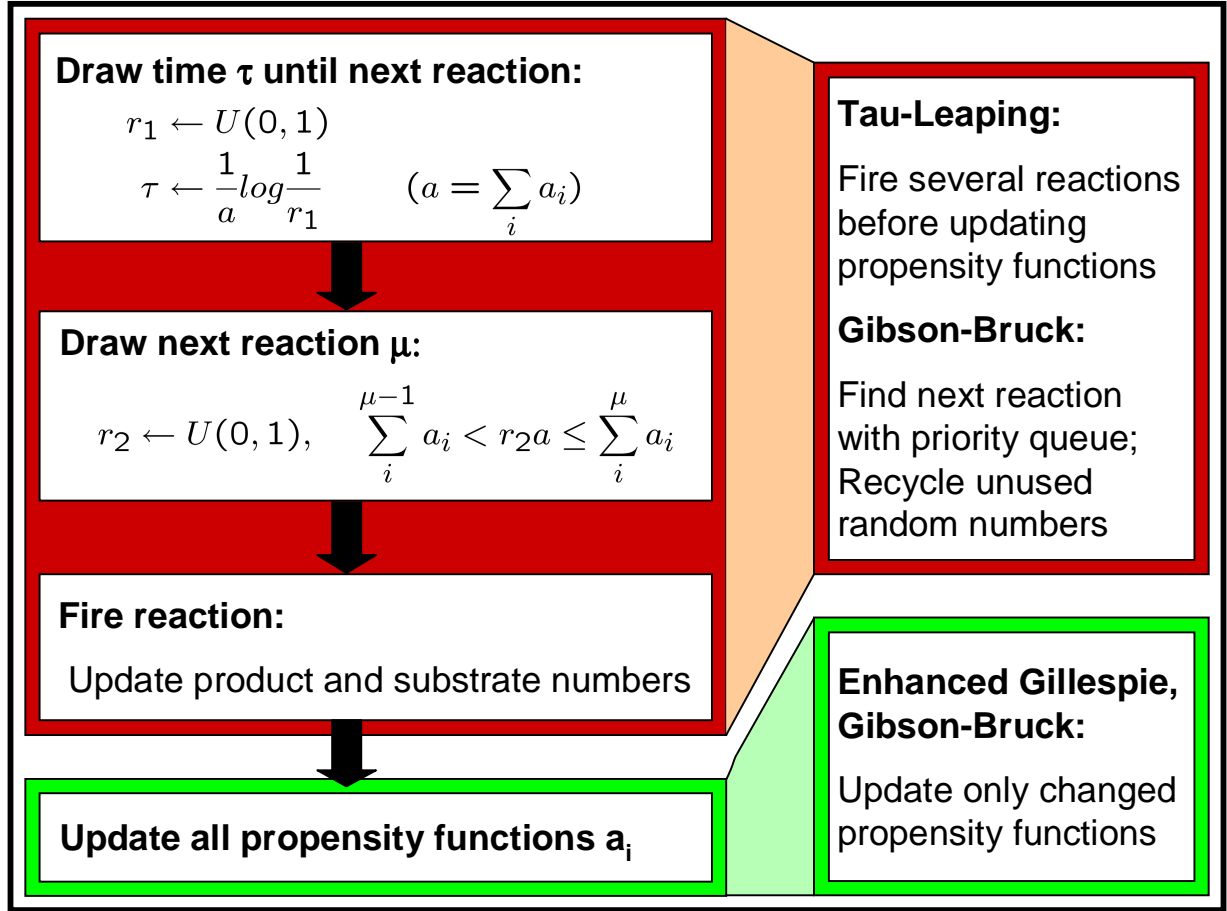


Figure 8.2: This figure shows the flow of one simulation step. On the left-hand side the flow for the original Gillespie algorithm can be seen. On the right-hand side, we illustrate how the different steps are modified by the Gibson-Bruck, enhanced Gillespie and tau-leaping algorithms. Here, $U(0, 1)$ denotes the uniform distribution on the range of 0 to 1 and a_μ the reaction propensity for reaction μ .

generate a uniformly distributed number u_i between 0 and 1 for each reaction and calculate $\tau_i = a_i(S)^{-1} \ln(u_i^{-1})$. Then the first reaction $i = \operatorname{argmin}_j \tau_j$ is determined and (τ_i, i) is taken as next pair (for a proof that this generates trajectories according to the CME, see [Gillespie, 1976]). Each firing thus requires generating $|T|$ exponentially distributed random numbers and a linear search to determine the first reaction.

Direct method

In order to draw the pair (τ, j) by the direct method [Gillespie, 1976], the joint probability (8.3) is rewritten to

$$p_i(\tau, j|S, t) = p'_i(\tau|S, t) \times p''_i(j|\tau, S, t) \quad (8.4)$$

$$p'_i(\tau|S, t) = a(S) \exp(-a(S)\tau) \quad (8.5)$$

$$p''_i(j|\tau, S, t) = \frac{a_j(S)}{a(S)} \quad (8.6)$$

where $a(S) = \sum_{j=1}^{|T|} a_j(S)$ is defined as above. Hence, τ is exponentially distributed with parameter $a(S)$, while j is distributed according to the discrete probabilities $\frac{a_j(S)}{a(S)}$ and independent of τ . The reaction j can be determined by generating a uniformly distributed random number u between 0 and 1 and then computing the smallest j such that $\sum_{j'=1}^j a_{j'}(S) > ua(S)$. Thus, each firing requires the generation of a single exponentially distributed random number and one uniformly distributed number, which is a substantial improvement compared to the first reaction method. Efficiently finding the smallest j is crucial to the performance of this method. Gillespie originally proposed a linear search method, which was later improved by [Cao et al., 2004] to a binary search method using a reordering of the propensities.

Next reaction method

24 years after the publication of the original SSA, Gibson and Bruck proposed some clever improvements of the first reaction method. It uses a priority queue to reuse previously unused tentative reaction times and a dependency graph for efficiently updating only those propensity functions, that have changed since the last firing. By using these additional data structures, it is possible to (asymptotically) generate only a single exponentially distributed number per firing [Gibson and Bruck, 2000]. However, it has been suggested that the next reaction method is actually less efficient than improved versions of the direct method [Cao et al., 2004] due to the cost of maintaining the data structures.

Composition/Rejection method

Generating a random number from a discrete probability distribution is not only a problem for SSAs but have also been encountered in other fields. Slepoy et al. [Slepoy et al., 2008] used a method previously described in [Devroye, 1986] to improve the direct method. If the minimum and maximum of the propensity functions is bounded (which is a valid assumption in biological networks), it is possible to generate j in expected constant time. This means that finding the next firing reaction does not depend on the number of reactions anymore. If the reaction network is sparse (not too many propensity functions have to be updated after firing a reaction), it becomes possible to simulate networks with tens of thousands of reactions.

Tau-leaping methods

The direct and next-reaction methods are exact methods. This means reaction propensities are updated after each reaction. Recently, Gillespie [Gillespie, 2001] proposed an approximative method, *tau-leaping*, which performs all reactions in a certain interval τ before updating the propensity functions (they 'leap' over a time interval). The interval size τ is chosen such that the propensity functions remain almost constant in this interval and reactions may fire multiple times. This, however, can sometimes lead to negative populations and as a consequence, this method has been improved later by Cao et al. [Cao et al., 2005a, 2006] to avoid this problem. The modified tau-leaping algorithm automatically switches to the exact SSA for a few steps if the choice of τ becomes too small. This allows for efficient simulation, especially when huge quantities of molecules are present in the system. If the interval is chosen such that the propensity function remain almost constant, the CME can accurately be approximated by the tau-leaping method.

Hybrid methods

Both exact and tau-leaping methods cannot be used to efficiently simulate models with multiple scales in molecule numbers or reaction rates. Exact methods are too inefficient to simulate many fast reactions and high molecule concentrations. On the other hand, the presence of low molecule concentrations and slow reactions in the systems will effectively lead to small τ values for the tau-leaping methods and thus make them behave as the exact methods. To circumvent these problems, hybrid methods have been developed, which partition the system into fast and slow reactions. The slow reactions are then generally simulated using the exact SSA. The fast reactions are solved either deterministically or with the Langevin equation [Cao et al., 2005b; Chiam et al., 2006; Harris and Clancy, 2006] or simulated with tau-leaping methods [Harris and Clancy, 2006; Puchalka and Kierzek, 2004]. Alternatively, the model is simplified such that the effect of the fast reactions is incorporated in the simulation of the slow reactions, e.g. using quasi-steady-state assumptions, without actually firing the fast reactions [Rao and Arkin, 2003; Cao et al., 2005b; Salis and Kaznessis, 2005; Cao et al., 2005c; Goutsias, 2005; Samant and Vlachos, 2005; Samant et al., 2007].

8.3 Implementation

The early implementations of SSAs (e.g. the one Gillespie proposed in [Gillespie, 1976]) in the late 70's were quite inflexible, since they used hardcoded reaction networks, i.e. the source code itself contained the definitions of S and $V = \{v_1, \dots, v_n\}$ and it was not possible to change it without editing the source code and recompiling it. With the advent of modern programming languages like C++ or Java, more flexible implementations have been developed.

Flexible in this context has several meanings. First, it is absolutely necessary to allow loading of reaction networks, i.e. the implementation should be able to load a reaction network stored in a file and simulate it. Second, the user should be able to observe or even control the simulation process in various ways. As indicated above, there are different methods for different fields of applications and using the best suited algorithm is crucial for simulation efficiency. As a

consequence, implementations have to include the various algorithms and should provide an easy way to add new methods, since even 32 years after the first SSA paper, there is still room for improvement [Slepoy et al., 2008].

8.3.1 Other implementations

Several implementations of stochastic simulation algorithms are available, e.g. COPASI [Hoops et al., 2006], Dizzy [Ramsey et al., 2005] using the SSA implementations of the ISBJava library, gillespie2 [Gillespie et al., 2006], STOCKS [Kierzek, 2002], StochKit [Li et al., 2007], and BioNetS [Adalsteinsson et al., 2004]. In general, these programs were designed as stand-alone programs and, as a consequence, the user is limited to the functionalities of the user interface. This makes it difficult to use the implementations of the SSAs within other programs. Furthermore, most of these programs provide only one implementation of an exact SSA method, which is not always fast enough for the specific needs of practical systems biology applications. Users cannot easily add faster SSAs such as e.g. the approximative tau-leaping procedure or new hybrid algorithms to the programs.

The StochKit software and ISBJava library provide these faster tau-leaping algorithms and the latter was also designed to be used within other systems biology programs. The output of the corresponding SSA implementations, however, is limited to the molecule concentrations. More flexible implementations are necessary to simulate complex high-level models and integrate stochastic simulation algorithms in new and innovative analysis and modeling tools. Two examples which illustrate the need for more flexible tools are the visualization of the simulation progress directly in a network and the simulation of cell growth and division. With current simulation tools, it is not possible to implement these two examples without having to change the code of the actual simulation algorithms considerably.

8.3.2 FERN

FERN [Erhard et al., 2008] is a Java framework for modeling and simulating biological systems, which provides all types of state-of-the-art simulation algorithms (exact, approximate and hybrid) and has been designed to be easily extendable to new ones. With the help of the observer programming pattern [Gamma et al., 2004], the simulation progress can be monitored on every level and modifications to the systems can be introduced during simulations in an intuitive way. Even with these additional functionalities, the implementation is faster than the respective ISBJava implementation. Results can be visualized easily and networks can be loaded from different sources. Contrary to ISBJava, FERN supports the most current version of SBML, a widely used file format for exchanging biological networks [Hucka et al., 2003], and allows arbitrary rate law definitions. FERN is primarily intended as a library which can be included in other Java applications to simulate reaction networks. However, various user interfaces are also provided by the FERN distribution, which allow to use it without writing a single line of Java code. The most basic user interface included in the FERN distribution is a command-line tool, which can be used to generate time courses for given species in a given reaction network from the unix shell or windows terminal. Additionally, by exploiting the plugin architectures of the

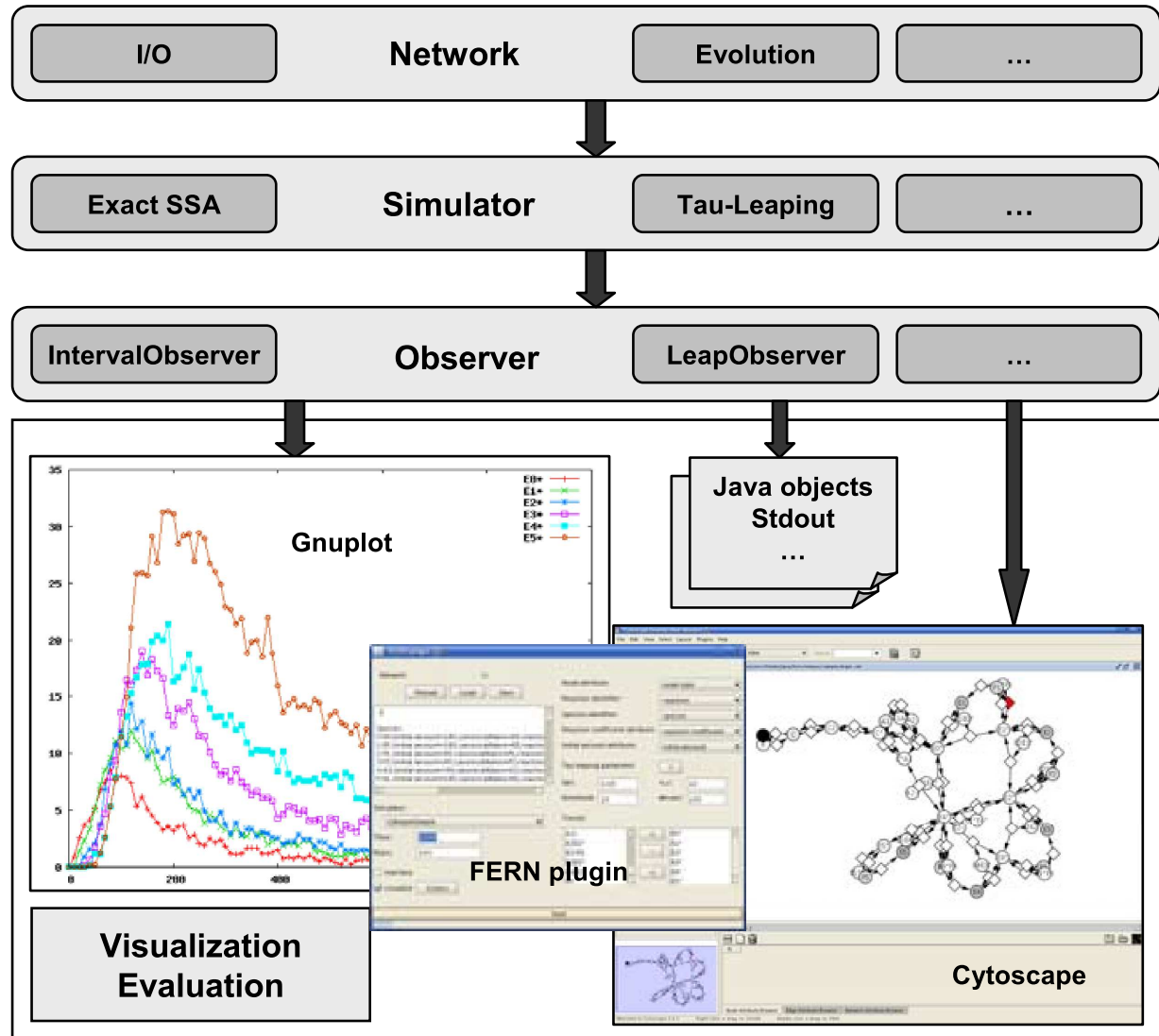


Figure 8.3: The figure illustrates the overall design of FERN into three layers. Each layer is represented by one interface or abstract class: Interface *Network* and abstract classes *Simulator* and *Observer*.

systems biology programs Cytoscape [Shannon et al., 2003] and CellDesigner [Funahashi et al., 2003], graphical interfaces to FERN are provided as well, which make it possible for the user to design reaction networks visually, simulate them, create time courses and even to observe the simulation process in real time on the graphical representation.

8.3.3 Implementation details

FERN is an object oriented library implemented in Java (see UML diagram in Figure 8.3). Although it consists of more than 100 classes and interfaces, most classes are implementations of one of three major interfaces and abstract classes:

1. The interface *Network* provides the network structure of the model.
2. The abstract class *Simulator* performs simulations on a *Network*. It additionally calls the registered observers during the simulation run.
3. The abstract class *Observer* traces the simulation progress and creates the simulation output.

A simple simulation can be performed in only five lines of code, one line for each of: loading a network file, creating a simulator, creating and registering an observer, running the simulations and printing the results. More complex examples for using FERN can be found in the FERN distribution. In the following, the three layers of FERN are described in more detail.

Networks

The interface *Network* describes the network's structure, i.e. the reactions and species in the networks. Furthermore, the network stores basic information like species names and their initial molecule numbers. For the simulation more information is necessary, which is stored in three additional classes:

- The *AmountManager* controls the amount of each molecular species during the course of a simulation.
- The *AnnotationManager* can store additional annotations for the network, its species and reactions.
- The *PropensityCalculator* calculates the propensity functions for the reactions by the specified kinetic laws.

There are three types of implementations of the *Network* interface:

- *Readers* which can read network data from files (e.g. FernMLNetwork, SBMLNetwork)
- *Decorators* which redirect method calls to existing network classes (e.g. CytoscapeNetworkWrapper)
- *Evolution algorithms* which generate networks according to certain network models (e.g. AutocatalyticNetwork)

For each network, stochastic simulations can be performed with all implemented simulation algorithms using all implemented observers.

An interesting example of network decorators and a perfect showcase for the flexibility of the FERN architecture are the modification networks. These are implementations of the *Network* interface, take an existing network instance and create a modified version of it. For instance, the *ReversibleNetwork* class creates a reverse reaction for each reaction of the input network. Here, the power of object oriented software design is exploited in two ways: First, the *ReversibleNetwork* can take any instance implementing the *Network* interface, e.g. a network loaded from the file system (*FernMLNetwork*), a network obtained from Cytoscape (*CytoscapeNetworkWrapper*) or even another modifier network. Second, the modifier network itself is an implementation of *Network* and can hence be used like every other network, e.g. for simulation. This scheme is often called *decorator pattern* [Gamma et al., 2004].

Import and export of networks

FERN supports two formats for loading and exporting networks: the SBML format [Hucka et al., 2003] as well as the simpler but also XML based FernML format. For reading and writing the SBML format, FERN uses the Java bindings of the C library (libSBML) available at <http://www.sbml.org>. Thus, it can be easily adapted to new developments of the SBML format.

From the model loaded by libSBML from the SBML file, a FERN *SBMLNetwork* is created using the list of compartments, species, reactions, parameters and events in the model. Events have to be registered with a simulator by the *SBMLNetwork* if they are to be triggered during the simulation. Triggering of events is handled by specific observers.

Currently, the *SBMLNetwork* class uses only the features of SBML necessary for the simulation of the network. It supports MathML to define complex reaction mechanisms but not rules, constraints or function definitions. If these features are required they can be incorporated easily by extending the *SBMLNetwork* class and loading these features from the SBML model created by libSBML. Since many systems biology applications support SBML (e.g. CellDesigner [Funahashi et al., 2003]), the SBML format can be used as an interchange format between FERN and these other applications.

SBML is a powerful format which can provide lots of information about a model. In contrast, FernML stores only the topology of the reaction network, optional annotations and the simulation parameters. This results in a much more simplified input format. More complex aspects, such as volume change due to cell growth and division, can then be modeled in Java using the FERN library in a straightforward way. As a consequence, arbitrarily complex models can be designed on the level of Java code.

Since FernML supports only the mass action reaction rate equations used by Gillespie [Gillespie, 1976], the propensities can be recalculated at each step efficiently by a few arithmetic operations. SBML uses MathML to store the kinetics of a reaction. This allows for more complex reaction mechanisms and is particularly useful if the model cannot be formulated exclusively with first or higher order rate equations. To evaluate MathML expressions, FERN creates expression trees from them, which have to be evaluated every time a propensity is calculated. Since this is one of the essential steps of SSAs, the simulation of an SBML network in FERN can be significantly

slower than the simulation of the same network as a FernML network. Thus, if only simple reaction rate equations are used, an SBML network should be converted to a FernML network using the provided conversion methods before performing the simulation.

FERN is not restricted to the input formats currently available. Any new input format can be easily included by implementing the *Network* interface or extending the *AbstractNetworkImpl* class.

Simulation algorithms

FERN provides implementations for four exact stochastic simulation algorithms, three state-of-the-art tau leaping procedures (see [Gillespie, 2001; Cao et al., 2006]) and a hybrid method combining SSA and tau-leaping [Puchalka and Kierzek, 2004]. The exact SSAs implemented include the original direct method of Gillespie [Gillespie, 1976], the next reaction method of Gibson and Bruck [Gibson and Bruck, 2000], the Composition/Reaction method [Slepoy et al., 2008] and an enhanced version of the direct method. This enhanced method uses the dependency graph technique of the next reaction method to only update the propensity functions that are affected by the firing of a reaction. Apart from this improvement, it is identical to the direct method. As indicated in [Cao et al., 2004], this simple enhancement is sufficient to make the direct method faster than the next reaction method in most cases.

The tau-leaping algorithms are all based on the modified tau-leaping procedure proposed by Cao et al. [Cao et al., 2005a], which avoids the problem of negative populations observed for the original tau-leaping procedure. This method switches to an exact SSA (in our implementations the enhanced Gillespie) for a few steps if the selected τ becomes too small. The three implementations differ only in the way the error is bounded (see [Cao et al., 2006] for details). The error is bounded either by the sum of all propensity functions (*TauLeapingAbsoluteBoundSimulator*), the relative changes in the individual propensity functions (*TauLeapingRelativeBoundSimulator*) or the relative changes in the molecular populations (*TauLeapingSpeciesPopulationBoundSimulator*).

Furthermore, FERN implements the hybrid method by Puchalka and Kierzek [Puchalka and Kierzek, 2004], which partitions the system during the simulation into slow reactions, which involve only small molecule numbers, and fast reactions, which involve large molecule numbers. The slow reactions are then simulated using an exact SSA while the fast reactions are simulated with tau-leaping. This algorithm was chosen over other hybrid methods for two reasons. First, it uses only stochastic simulation algorithms, i.e. exact SSA and tau-leaping, and no further assumptions such as quasi-steady state. Second, the partitioning of the system is performed dynamically according to the state of the system and updated after each reaction step. Our implementation of the hybrid method uses our more efficient enhanced Gillespie algorithm instead of the Gibson and Bruck algorithm used by Puchalka and Kierzek. On the model of LacZ and LacY gene expression by Kierzek [Kierzek, 2002], the hybrid method speeds up the runtime by a factor of about 100 compared to the enhanced Gillespie algorithm.

Future developments of the algorithms can easily be included into FERN by extending one of the SSA implementations or the original *Simulator* class. In the same way, ODE solvers or simulators for spatial models, which are not provided by FERN, can be integrated.

Observer system

FERN uses observers [Gamma et al., 2004] to trace the simulation progress and react to events. For this purpose, each observer has to implement functions which describe its response at specific time points of the simulations. Such responses may occur either at the beginning or the end of a simulation, before each step, after a reaction is fired or when a certain time is reached. In order to be notified of these events, observers have to be registered with the simulator.

Observer implementations are provided for tracing the molecule numbers for some species in arbitrary intervals, for recording the firings of reactions, for computing distributions of molecule numbers at a certain time over many simulation runs as well as for many other purposes. Several observers can be registered for a simulation at the same time and most of them can also handle repeated simulation runs, e.g. to create average curves or curves containing all trajectories for the individual simulation runs.

In general, the observers use gnuplot to present their results. Once gnuplot is installed on a system and accessible e.g. via the path variable, the *Gnuplot* class makes it possible to easily create plots and retrieve them as *Image* objects, save them as files or present and update them online in a window. Plots can be customized using appropriate gnuplot commands.

Stochastics

An important feature of FERN is that random number generation is handled by the singleton class *Stochastics*. Accordingly, only one instance of this class is instantiated during a FERN run and all calls for random numbers are referred to this instance. This has several advantages. First, the underlying random number generator can easily be replaced if faster and better random number generators are developed. Currently, the Mersenne Twister implementation of the Colt Project is used (<http://dsd.lbl.gov/~hoschek/colt/>). Second, by setting the seed value for the random number generator explicitly, the simulation can be made deterministic and e.g. interesting trajectories can be reproduced. Third, it is possible to count the number of random number generations necessary for different implementations of SSAs, which is particularly interesting to figure out why some algorithm is inefficient for some application.

8.3.4 Accuracy and runtime performance of FERN

To test the accuracy of the implemented stochastic simulation algorithms we used the Discrete Stochastic Models Test Suite (DSMTS) [Evans et al., 2007]. This test suite provides 36 stochastic models in the SBML format which have been solved either analytically or numerically. To test the implementation of a stochastic simulation algorithm, simulations have to be performed a large number of times (in general 10,000 times) for each individual model. The test is failed for a model if the distribution of the results is statistically significantly different from the known underlying distribution.

The exact simulation algorithms passed 94.4% of the DSMTS models [Erhard et al., 2008], which is significantly better than the performance of other implementations. The reference

implementation of one of the authors of DSMTS, gillespie2 [Gillespie et al., 2006] for instance passes only 80.6% of the tests.

Even though FERN is implemented in Java, which is often claimed to be less efficient than C, FERN's original gillespie algorithm is significantly faster than the C implementation of gillespie2. FERN was also compared to the SSA implementation of ISBJava using the EGF signaling pathway by Lee et al. [Lee et al., 2006], which contains 39 molecular species and 19 reversible and 12 irreversible reactions. Our results show that the implementations of the original Gillespie and Gibson-Bruck algorithm of FERN are always more efficient for the FernML network than the implementations provided by ISBJava (for details about runtime performance, see to [Erhard et al., 2008]).

Furthermore, the enhanced implementation of the Gillespie algorithm provided by FERN outperforms any of the exact methods provided by ISBJava. This shows that the powerful observer system of FERN does not come at the cost of a reduced runtime performance. In contrast, observers may rather help to avoid the execution of unnecessary code. Accordingly, FERN is a useful library for stochastic simulation even if the observer tools are not used.

8.4 Using FERN

8.4.1 Command line tool

FERN can directly be used from command line by calling one of the start scripts. This is currently supported for windows (start.bat), linux/unix/mac os (start.sh) and cygwin (start_cygwin.sh). They all call the main method of the class *fern.Start*, which is able to simulate networks with different algorithms and record time courses of molecular species. Usually, these time courses are written to stdout in tab separated format, to enable an easy processing by downstream programs. In addition, the command line tool of FERN is able to directly create plots of time courses using the freely available software gnuplot.

In the following, some examples are shown for using the command line tool to simulate the EGF signaling pathway by Lee et al. [Lee et al., 2006], which is included in the FERN distribution as a FernML network (examples/mapk.xml) and an SBML network (examples/mapk_sbml.xml).

The most basic simulation run can be started by typing

```
start.sh examples/mapk.xml 800 10
```

This will simulate the pathway for a time of 800 seconds and record the amount of each of the 39 molecular species every 10 seconds. The recorded time course is written to stdout, which gives a total of 81 lines of 40 columns each containing the time in the first column and the molecule number of the *i*'th molecule in the *i*'th column (the molecular species are numbered according to their occurrence in the xml file):

```
0.0      680.0    100.0    0.0      0.0      0.0      ...
...
800.0    580.0     0.0     2.0     0.0     12.0     ...
```


In the model, the molecular species $E0^*$, $E1^*$, $E2^*$, $E3^*$, $E4^*$, $E5^*$ represent the receptor signaling complex EGF.EGFR2Grb2.SOS ($E0^*$) and the phosphorylated forms of the kinases Ras, Raf, MEK, ERK and Elk ($E1^* - E5^*$). Since these are the constituent parts of the phosphorylation cascade and therefore particularly interesting, it is possible to specify the molecular species to observe using the `-s` parameter.

The output then will only have seven columns: the time and the amount of these six molecular species (now the species are numbered according to their occurrence in the command). In order to create a plot, the *interactive* flag can be used.

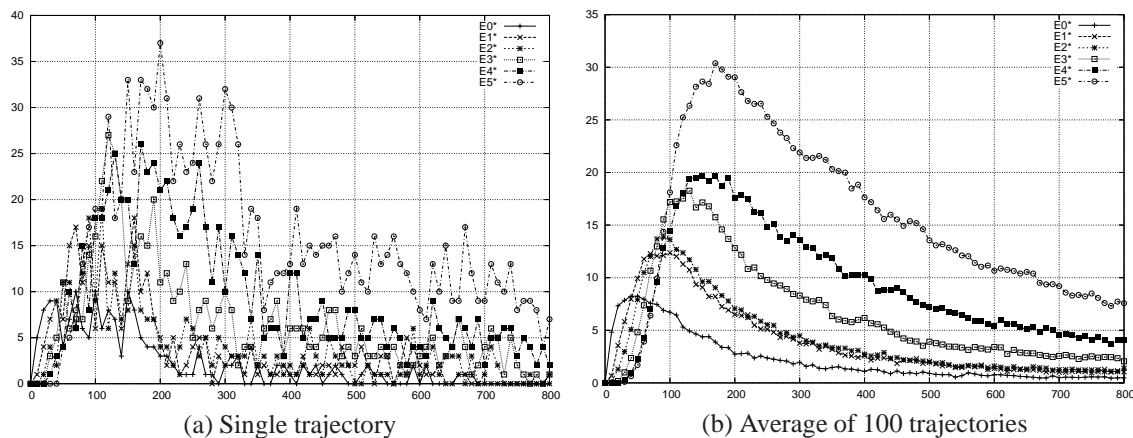


Figure 8.4: Trajectories of the EGF signalling pathway by Lee et al. [Lee et al., 2006]. After 100 SSA runs, the average of all trajectories strongly resembles the time course of the ODE model.

A window containing the time course plot (see Figure 8.4) will pop up once the simulation has finished. It is possible to save the plot in various formats by clicking onto the window. Lee et al. used ODEs to model the pathway, so their plots looked a lot smoother than the time course created by SSAs, which is a random sample according to the CME. In order to create multiple time courses, use the `-n` parameter. The full command then is:

```
start.sh examples/mapk.xml 800 10 -i -n 100 \
-s E0* E1* E2* E3* E4* E5*
```

Now, the plot is updated after every simulation run. The more runs are averaged, the more similar are the results to the ODE model in [Lee et al., 2006] as the CME is better approximated.

8.4.2 Basic usage of FERN

Apart from the the command line, Java programmers can use FERN's advanced functionalities by writing only a few lines of code. A network can be loaded in a single line of Java code (line 1 in Listing 8.2). Then it can be simulated using a *Simulator* instance (lines 4 and 9).

The parameter 800.0 is the time duration for simulation, i.e. the reaction network is simulated from $t_0 = 0$ with S_0 = the initial amounts taken from mapk.xml to time $t_N = 800$. Simulation

results, however, are not reported yet (and not even recorded), since no *Observer* is registered at the simulator. The *AmountIntervalObserver* can be used to create time courses of molecular populations (lines 5 and 6).

When the simulation is started afterwards, the observer records the amounts of the phosphorylated kinases (see above) at time points $t_0 = 0, t_{i_1} = 10, t_{i_2} = 20, \dots, t_{i_{81}} = 800$. If the line *sim.start(800.0)* is placed inside a for loop (line 8), the observer will record all the runs. However, these results are still not reported to the user.

This is done by lines 11-16. These create a plot of the averaged time courses as well as the tab separated output known from the command line tool.

```

1 Network net = new FernMLNetwork(new File("mapk.xml"));
2 NetworkTools.dumpNetwork(net, new PrintWriter(System.out));
3
4 Simulator sim = new GillespieEnhanced(net);
5 AmountIntervalObserver amount = new AmountIntervalObserver(sim,10,"
  E0*", "E1*", "E2*", "E3*", "E4*", "E5*");
6 sim.addObserver(amount);
7
8 for (int i=0; i<100; i++)
9   sim.start(800.0);
10
11 GnuPlot gp = new GnuPlot();
12 gp.setVisible(true);
13 amount.toGnuplot(gp);
14 gp.plot();
15
16 System.out.println(gp.toString());

```

Listing 8.2: Complete listing of the example

Although this example only performs the same simulation as the command line tool, it illustrates the structure and modularity of a FERN program for the use of even inexperienced Java programmers.

Additional examples are included in the FERN distribution in the package *fern.example*, which can be used as a starting point for own projects. Besides these single class examples, the FERN distribution includes more sophisticated applications of its functionalities, which are described in the following.

8.4.3 Cytoscape plugin for stochastic simulation

Cytoscape [Shannon et al., 2003] is a software platform for visualizing and integrating networks with an emphasis on biological data. It provides a flexible plugin architecture, which can be used to enrich the platform with additional methods. We used this functionality to create a plugin which uses FERN to simulate networks loaded into Cytoscape (see Figure 8.5). This plugin

makes it possible to track the simulation progress directly on the network. Furthermore, it shows how FERN can be easily integrated into other applications and how the observer system can be used to visualize more than just the changes in molecule numbers.

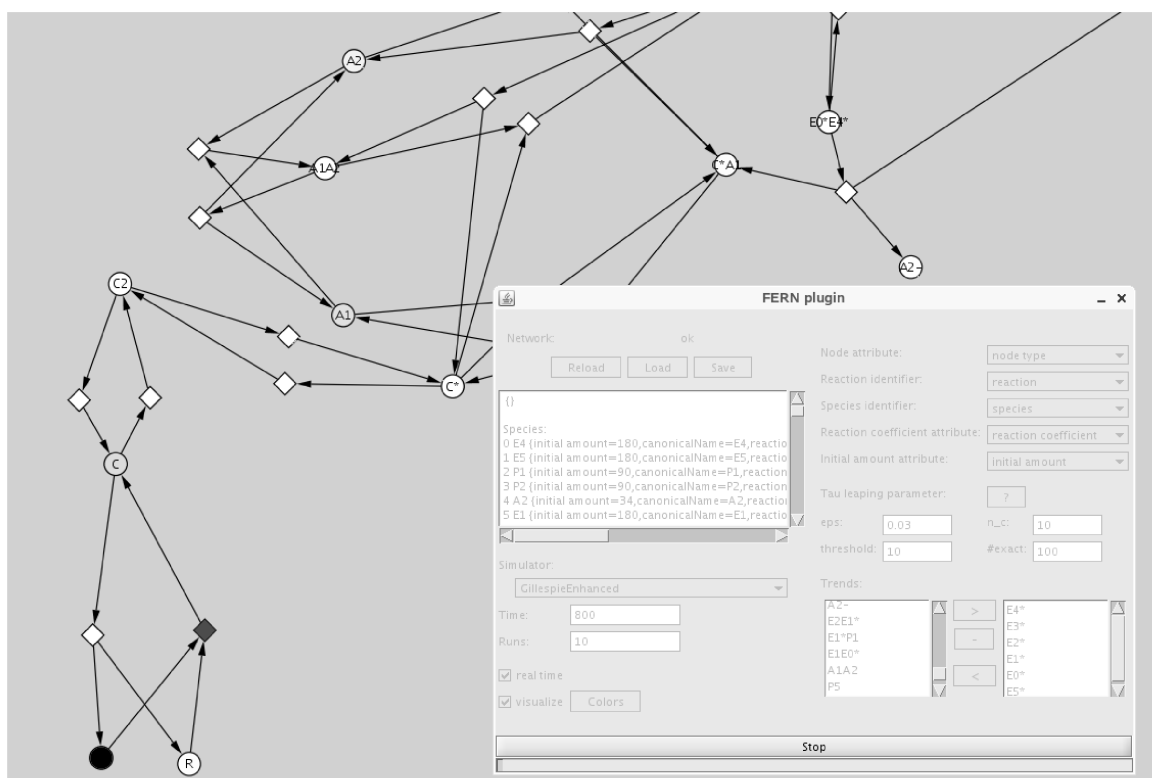


Figure 8.5: EGF signalling pathway loaded into Cytoscape. Currently, a simulation is performed and visualized directly on the graphical representation of the network: The transition $L + R \rightarrow C$ (the dark colored diamond in the lower left corner) is currently firing.

Each network readable by Cytoscape can be used for simulation by the plugin if it consists of two distinct types of nodes, namely reactions and molecular species. Furthermore, the initial amount of each molecular species and the reaction rate coefficient for each reaction are required. These parameters and the node type (species or reaction) can be read from arbitrary node attributes specified in Cytoscape. Additionally, the plugin provides access to FernML files in both directions. Thus every Cytoscape network can be saved as FernML, and every FernML file can be loaded into Cytoscape.

Simulations can be performed with every stochastic simulation algorithm provided by FERN and the simulation progress can be visualized directly on the network. Reaction nodes flash up whenever the corresponding reaction is fired and the species nodes are colored according to their molecule numbers. Furthermore, simulations can be run in real-time, which causes the algorithms

to pause between two reaction events according to the simulated time. In addition, time series of molecular species can be created using gnuplot.

The implementation of the Cytoscape plugin is straightforward. A central plugin class integrates FERN into the Cytoscape platform by creating a menu item to start the plugin and to load the user interface. Apart from the classes defining the user interface, only a few additional classes are necessary. The most important ones are a wrapper class implementing the *Network* interface to map the Cytoscape network structure to FERN and an *Observer* class to make the visualization possible. Additionally, FERN provides its own *Visual Style* (which defines how nodes and edges are colored and shaped in Cytoscape) to guarantee a proper display of the network and to handle the flashing and recoloring of reaction and species nodes, respectively.

The Cytoscape plugin was also adapted as a plugin to CellDesigner [Funahashi et al., 2003], which now offers a plugin functionality with the recent version 4.0 beta.

8.4.4 Simulation of cell growth and division using observers

The Cytoscape plugin is one example how observers can be used to track the simulation progress at various levels. Another example which illustrates the potential of the observer system is the simulation of the LacZ model described by Kierzek et al. [Kierzek et al., 2001; Kierzek, 2002] and based on experimental results by Kennell and Riezman [Kennell and Riezman, 1977].

This model requires the simulation of cell division. After each cell division, the stochastic simulation is continued with one promotor molecule and all other molecule numbers divided by 2. RNA polymerase and ribosome molecules are assumed to remain approximately constant with natural variations. For this purpose, the number of these molecules has to be adjusted after each simulation step by drawing from normal distributions. Furthermore, cell growth leads to a linear volume change.

With existing stochastic simulation programs, this model can, in general, only be simulated by changing the code of the actual simulation algorithms. Contrary to that, the model can be easily simulated with FERN by simply defining a cell growth observer. Before each simulation step, the observer checks if a generation has been completed. If this is the case, all molecule numbers are adjusted as described before. In any case, the volume size is adjusted to account for either cell division or cell growth, and the RNA polymerase and ribosome molecule numbers are set randomly.

This approximation was also used by Kierzek et al. and assumes that cell volume does not change during a simulation step. To perform an exact simulation of volume change, propensity functions would have to be defined which handle the cell volume as a function of time. However, since the volume change during one reaction is extremely small, the differences between the approximate and exact results should be negligible.

Using the cell growth observer, we simulated the LacZ model with the enhanced Gillespie algorithm. Our results for the concentration of the LacZ protein clearly show the periodic oscillation in the protein numbers due to cell growth and division (see Figure 8.6). From these results, we can estimate the rate of LacZ protein synthesis via a linear fit to the increasing LacZ concentrations during the first generation. Here, we obtained a rate of protein synthesis of $21 s^{-1}$,

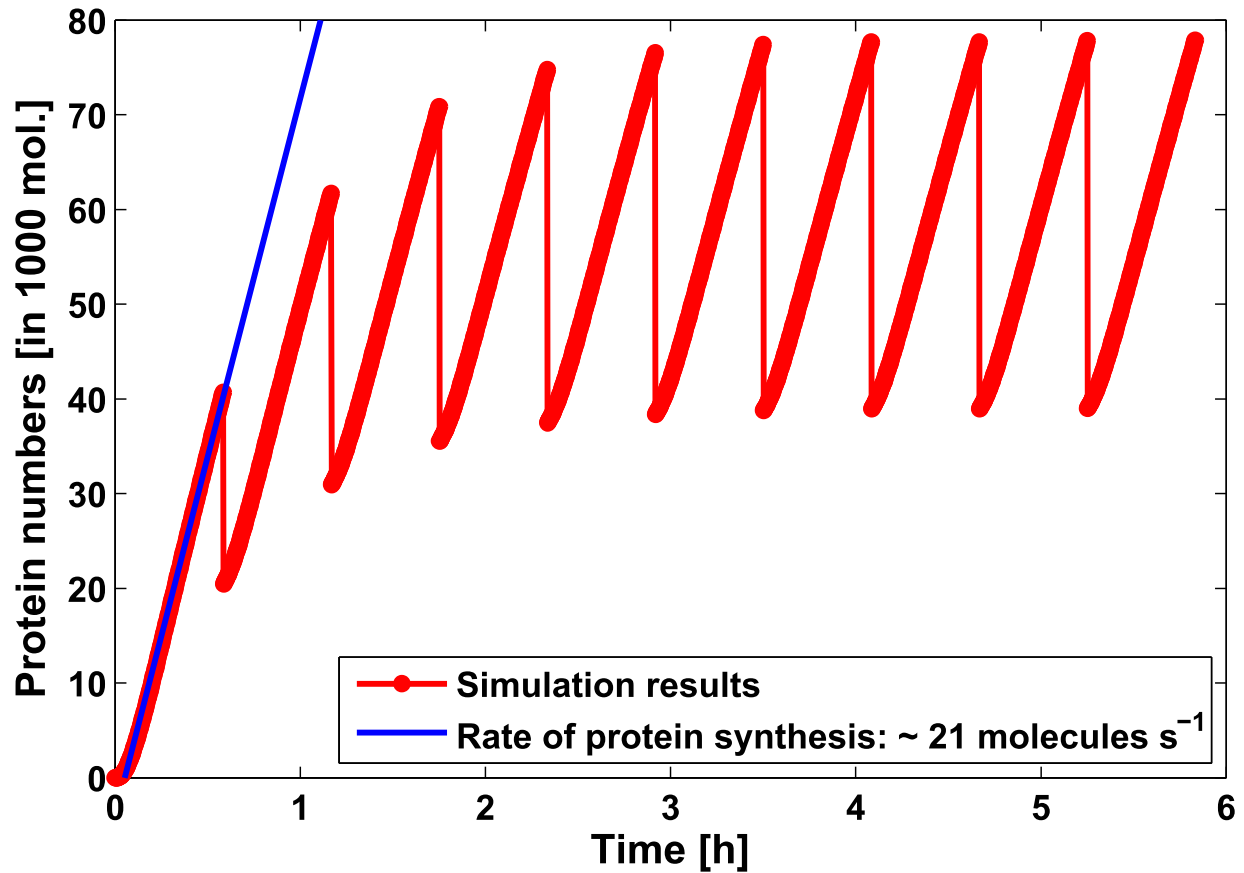


Figure 8.6: Average results of 1,000 simulations are shown for the LacZ protein over ten bacterial generations (black). After each generation (35 min) the number molecules for each species was divided by 2 to simulate cell division. The gray line shows a linear fit to the increasing LacZ concentration during the first generation. This yields a rate of protein synthesis of $21s^{-1}$.

which is close to the $22s^{-1}$ obtained by Kierzek et al. [Kierzek, 2002] and the $20s^{-1}$ reported by Kennell and Riezman [Kennell and Riezman, 1977].

8.5 Discussion

In this chapter, we presented FERN, a Java framework for modeling and simulating biological reaction networks. We showed, that FERN improves on implementations in terms of runtime efficiency and flexibility and that it provides a comprehensive and easy to use framework for fast, accurate and flexible stochastic simulation to Java developers. It provides state-of-the-art stochastic simulation algorithms, efficient representations of networks with several input and output options and various ways of tracing and visualizing simulation data.

It is possible to do reasonable simulations with FERN in just five lines of Java code. Each of the five steps can be expanded to cover more complex scenarios and simulations can be controlled

at different levels. For instance, to simulate cell growth, an observer can be modified to change the volume of the simulation space. Alternatively, an interesting subnetwork can be selected on which simulations can then be run.

Compared with the ISBJava library, FERN has several advantages. First, FERN is more flexible than ISBJava and offers more functions for tracking and interacting with simulations. Second, in contrast to ISBJava, it implements the composition/rejection method for large networks, a hybrid algorithm as well as the most current tau-leaping methods which resolve the problem of negative concentrations. Furthermore, its stochastic simulation algorithms are significantly faster than the ISBJava implementations. Finally, it supports the current version of SBML and allows arbitrary rate laws.

FERN can be easily integrated into other applications making its functionalities available within different environments. We have illustrated this by implementing FERN plugins to Cytoscape and CellDesigner. With only few additional classes, the Cytoscape plugin enables the users to follow the simulation progress directly on the network. This was made possible by the powerful observer system of FERN, which is one of its major advantages compared to other available simulation programs. In addition, we currently develop PNMA (Petri Net Modeling Application), a software platform for modeling, simulation and parameter optimization of biological networks based on Petri nets (<http://www.bio.ifi.lmu.de/PNMA>). It also includes a FERN plugin for stochastic simulation and offers many more specialized functionalities regarding Petri nets and their simulation than Cytoscape or CellDesigner.

Thus, the plugins and the command line tool make it possible to exploit FERN's functionalities without writing Java code. Although some available stochastic simulation programs offer a few specialized features not yet supported by FERN such as e.g. time-delayed dynamics, none of them offer such a wide range of features and can be extended to new features as easily as FERN. Therefore we provide FERN as a useful tool for biochemical network analysis or the development of new analysis methods or applications.

Chapter 9

Conclusion and outlook

In this work I developed several methods to analyze and interpret high-throughput data from systems biology experiments. As pointed out in the previous chapters for specific examples, it is highly important to properly analyze the raw data from such experiments. Otherwise, interesting findings may be missed or, which is even worse, mistakes made in the raw data analysis, e.g. not recognizing certain bias, may lead to wrong interpretations. Furthermore, it becomes more and more apparent that the major bottleneck in genomics research is not the generation of data but their computational analysis, interpretation, visualization and integration [Green and Guyer, 2011]. For instance, the ENCODE project [Consortium, 2012b] has generated and is still generating unprecedented and massive amounts of data. Even if various papers have been published jointly in *Nature*, *Genome Research* and *Genome Biology* in September 2012, there is probably still much to be found in the ENCODE data, and hypotheses published by the participants of the ENCODE project must be challenged. Only computational methods and tools give the opportunity to handle such data.

I applied the methods I developed to data that was obtained in the context of a project about herpes viruses with a focus on post-transcriptional regulation by microRNAs. ALPS (see chapter 3 and Erhard and Zimmer [2010]), PARma (see chapter 4 and Erhard et al. [2013a]) and REA (see chapter 5 and Erhard et al. [2013b]) are methods for the raw data analysis of short RNA-seq, PAR-CLIP and RIP-Chip experiments, respectively. All of these experiment types are widely used in biological research and implementations of ALPS, PARma and REA are publicly available on the institute's website and can be used by the research community to identify regulatory RNA in short RNA-seq experiments or to find their targets by PAR-CLIP or RIP-Chip experiments.

These methods or ideas thereof may also help to answer additional questions that have not been addressed so far. For instance, since regulatory ncRNAs are also sequenced in a PAR-CLIP experiment, it may be of interest to apply ALPS to PAR-CLIP data in order to identify novel regulatory ncRNAs that is also utilized by RISC to recognize targets. Furthermore, the ALPS scoring system could be extended to also consider PAR-CLIP characteristic features, i.e. T to C conversions.

Furthermore, ideas from PARma could also be used to analyze other types of experiments, for instance Digital genomic footprinting (DGF) [Neph et al., 2012b]. Roughly, in comparison to

CLIP-seq, which measures data of microRNA-mediated regulation, DGF can be seen as the equivalent experiment for data of transcription factor (TF)-mediated regulation. In DGF, binding sites of all active TFs are obtained in a genome-wide manner, which is similar to CLIP-seq experiments, where binding sites of microRNAs are measured in a genome-wide manner: In both cases, there are two tasks for bioinformatics, the identification of the binding sites and of the regulator that has bound to each of the sites. Importantly, while the characteristic features of PAR-CLIP data are uniform for each microRNA [Erhard et al., 2013a], DGF target sites may exhibit quite distinct but highly specific patterns in the DGF data depending on the TF that has bound there [Neph et al., 2012b]. Thus, PARma cannot be directly applied to DGF data, but extending PARma to respect these distinct patterns may help to distinguish TFs with high accuracy.

FERN (see chapter 8 and [Erhard et al., 2008]) is a software package for stochastic simulation of Petri nets. The simulation of biological networks will become more and more important in the future due to the massive amounts of data that is being produced and tools such as FERN will be important to model and check the measurements by simulation. FERN is also freely available on the internet as stand-alone library, as a plugin for Cytoscape and for CellDesigner, and also embedded in our in-house modelling applications PNMA.

The detection of outlier peptides (see chapter 7 and Erhard and Zimmer [2012]) is a first step towards finding differentially spliced genes in proteomics data. Considering transcript level expression instead of gene level expression is a necessary step towards understanding eukaryotic gene regulation and technological advances in mass spectrometry and their analysis methods will help to uncover splicing patterns on proteins in a genome-wide manner. Furthermore, ncRNAs have been implicated in the regulation of splicing and thus, integrating microRNA related experimental data, e.g. target sites obtained by PAR-CLIP, and experiments measuring splicing patterns will be an important topic in the future.

The most intriguing finding in this work is that microRNA-mediated regulation is dependent on the context and that context-dependence is not restricted to a few examples but a widespread feature for microRNAs. Future experiments and research must focus on the identification of contributors to this context. For instance, there are already several studies available that try to identify binding sites either of specific RNA binding proteins [Lebedeva et al., 2011] or of all RNA binding proteins at once [Baltz et al., 2012]. However, such experiments must be conducted for multiple contexts in conjunction with experiments that measure microRNA targets. Thus, not only an *Encyclopedia of DNA Elements* is necessary to understand the human genome, but also an *Encyclopedia of RNA Elements*.

So far, high-throughput data has mainly played a role in basic research. However, falling prices are currently initiating a new age for high-throughput experiments such as NGS: They are more and more applied in a clinical setting, opening up completely new opportunities for personalized medicine [Biesecker et al., 2012]. Thus, the development of analysis software that cannot only be used by specially trained bioinformaticians but also by clinical personal will also play an important role in the future.

Currently, we may not yet be in the position to understand a complete system such as a single cell or a whole organism in a quantitative manner, but high-throughput technologies give us the opportunity for big steps towards a quantitative understanding. It is in the responsibility of

computational biology to take these steps by providing methods and tools for the analysis and interpretation of high-throughput data.

Bibliography

- Adalsteinsson, D., McMillen, D., and Elston, T. C. Biochemical network stochastic simulator (bionets): software for stochastic modeling of biochemical networks. *BMC Bioinformatics*, 5:24, 2004.
- Ambros, V. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in *c. elegans*. *Cell*, 57(1):49–57, 1989.
- Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.
- Aravin, A. A., Naumova, N. M., Tulin, A. V., Vagin, V. V., Rozovsky, Y. M., and Gvozdev, V. A. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.*, 11:1017–1027, 2001.
- Arvey, A., Agius, P., Noble, W. S., and Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome research*, 22(9):1723–1734, 2012.
- Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P., and Blelloch, R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev.*, 22:2773–2785, 2008.
- Bachellerie, J. P., Cavaille, J., and Huttenhofer, A. The expanding snoRNA world. *Biochimie*, 84:775–790, 2002.
- Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, 2008.
- Baltz, A. G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., et al. The mRNA-Bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular Cell*, 46(5):674–690, 2012.
- Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.
- Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.

- Bazzini, A. A., Lee, M. T., and Giraldez, A. J. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science*, 336(6078):233–237, 2012.
- Beitzinger, M., Peters, L., Zhu, J. Y., Kremmer, E., and Meister, G. Identification of human microRNA targets from isolated argonaute protein complexes. *RNA Biology*, 4(2):76–84, 2007.
- Ben-Bassat, H., Goldblum, N., Mitrani, S., Goldblum, T., Yoffey, J. M., Cohen, M. M., Bentwich, Z., Ramot, B., Klein, E., and Klein, G. Establishment in continuous culture of a new type of lymphocyte from a "Burkitt like" malignant lymphoma (line D.G.-75). *International journal of cancer. Journal international du cancer*, 19(1):27–33, 1977.
- Ben-Moshe, N. B., Avraham, R., Kedmi, M., Zeisel, A., Yitzhaky, A., Yarden, Y., and Domany, E. Context-specific microRNA analysis: identification of functional microRNAs and their mRNA targets. *Nucleic Acids Research*, 40(21):10614–10627, 2012.
- Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
- Bentwich, I. Prediction and validation of microRNAs and their targets. *FEBS Lett.*, 579:5904–5910, 2005.
- Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R. H. A. Diversity of microRNAs in human and chimpanzee brain. *Nature genetics*, 38(12):1375–1377, 2006.
- Bertsch, A., Gröpl, C., Reinert, K., and Kohlbacher, O. OpenMS and TOPP: open source software for LC-MS data analysis. In M. Hamacher, M. Eisenacher, and C. Stephan, editors, *Data Mining in Proteomics*, volume 696, pages 353–367. Humana Press, Totowa, NJ, 2011. ISBN 978-1-60761-986-4.
- Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*, 11(8):R90, 2010.
- Bhattacharyya, S. N., Habermacher, R., Martine, U., Closs, E. I., and Filipowicz, W. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, 125(6):1111–1124, 2006.
- Biesecker, L. G., Burke, W., Kohane, I., Plon, S. E., and Zimmern, R. Next-generation sequencing in the clinic: are we ready? *Nature reviews. Genetics*, 13(11):818–824, 2012.
- Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.
- Bodenmiller, B., Campbell, D., Gerrits, B., Lam, H., Jovanovic, M., Picotti, P., Schlapbach, R., and Aebersold, R. PhosphoPep – a database of protein phosphorylation sites in model

- organisms. *Nat Biotech*, 26(12):1339–1340, 2008.
- Boger, H. P., Karnowski, H. W., Cai, X., Shin, J., Pohlers, M., and Cullen, B. R. A mammalian herpesvirus uses noncanonical expression and processing mechanisms to generate viral MicroRNAs. *Molecular cell*, 37(1):135–142, 2010.
- Boutz, P. L., Chawla, G., Stoilov, P., and Black, D. L. MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development. *Genes & Development*, 21(1):71–84, 2007.
- Breitling, R., Amtmann, A., and Herzyk, P. Iterative group analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC bioinformatics*, 5:34, 2004.
- Brennecke, J., Stark, A., Russell, R. B., and Cohen, S. M. Principles of MicroRNA Target recognition. *PLoS Biology*, 3(3):e85, 2005.
- Calin, G. A. and Croce, C. M. MicroRNA signatures in human cancers. *Nature Reviews Cancer*, 6(11):857–866, 2006.
- Calzone, L., Thieffry, D., Tyson, J. J., and Novak, B. Dynamical modeling of syncytial mitotic cycles in drosophila embryos. *Mol Syst Biol*, 3:131, 2007.
- Cao, Y., Gillespie, D. T., and Petzold, L. R. Avoiding negative populations in explicit poisson tau-leaping. *J Chem Phys*, 123(5):054104, 2005a.
- Cao, Y., Gillespie, D. T., and Petzold, L. R. Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *Journal of Computational Physics*, 206:395–411, 2005b.
- Cao, Y., Gillespie, D. T., and Petzold, L. R. The slow-scale stochastic simulation algorithm. *J Chem Phys*, 122(1):14116, 2005c.
- Cao, Y., Gillespie, D. T., and Petzold, L. R. Efficient step size selection for the tau-leaping simulation method. *Journal of Chemical Physics*, 124:044109, 2006.
- Cao, Y., Li, H., and Petzold, L. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *J Chem Phys*, 121(9):4059–4067, 2004.
- Cazalla, D., Xie, M., and Steitz, J. A. A primate herpesvirus uses the integrator complex to generate viral microRNAs. *Molecular cell*, 43(6):982–992, 2011.
- Cazalla, D., Yario, T., and Steitz, J. A. Down-regulation of a host MicroRNA by a herpesvirus saimiri noncoding RNA. *Science*, 328(5985):1563–1566, 2010.
- Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147(2):358–369, 2011.

- Chan, P. P. and Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, 37:D93–97, 2009.
- Chawla, G., Lin, C., Han, A., Shiue, L., Ares, M., and Black, D. L. Sam68 regulates a set of alternatively spliced exons during neurogenesis. *Mol. Cell. Biol.*, 29(1):201–213, 2009.
- Cheloufi, S., Dos Santos, C. O., Chong, M. M. W., and Hannon, G. J. A dicer-independent miRNA biogenesis pathway that requires ago catalysis. *Nature*, 465(7298):584–589, 2010.
- Chen, K. and Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, 8(2):93–103, 2007.
- Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, 2009.
- Chiam, K.-H., Tan, C. M., Bhargava, V., and Rajagopal, G. Hybrid simulations of stochastic reaction-diffusion processes for modeling intracellular signaling pathways. *Phys Rev E Stat Nonlin Soft Matter Phys*, 74(5 Pt 1):051910, 2006.
- Christodoulou, F., Raible, F., Tomer, R., Simakov, O., Trachana, K., Klaus, S., Snyman, H., Hannon, G. J., Bork, P., and Arendt, D. Ancient animal microRNAs and the evolution of tissue identity. *Nature*, 463(7284):1084–1088, 2010.
- Cifuentes, D., Xue, H., Taylor, D. W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G. J., Lawson, N. D., et al. A novel miRNA processing pathway independent of dicer requires argonaute2 catalytic activity. *Science*, 328(5986):1694–1698, 2010.
- Clodong, S., Dhring, U., Kronk, L., Wilde, A., Axmann, I., Herzog, H., and Kollmann, M. Functioning and robustness of a bacterial circadian clock. *Mol Syst Biol*, 3:90, 2007.
- Colinge, J. and Bennett, K. L. Introduction to computational proteomics. *PLoS Comput Biol*, 3(7):e114, 2007.
- Consortium, T. . G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012a.
- Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012b.
- Cooper, T. A. Alternative splicing regulation impacts heart development. *Cell*, 120(1):1–2, 2005.
- Cooper, T. A., Wan, L., and Dreyfuss, G. RNA and disease. *Cell*, 136(4):777–793, 2009.
- Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., and Ohler, U. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biology*, 12(8):R79, 2011.
- Cortina, J. M. and Nouri, H. *Effect size for ANOVA designs*. Quantitative Applications in the Social Sciences. SAGE University paper, 2000. ISBN 9780761915508.

- Cox, J. and Mann, M. Is proteomics the new genomics? *Cell*, 130(3):395–398, 2007.
- Cox, J. and Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotech*, 26(12):1367–1372, 2008.
- Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research*, 10(4):1794–1805, 2011.
- Craig, R. and Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.
- Cullen, B. R. Viruses and microRNAs. *Nature Genetics*, 38:S25–S30, 2006.
- Cullen, B. R. Five questions about viruses and MicroRNAs. *PLoS Pathog*, 6(2):e1000787, 2010.
- Cullen, B. R. Viruses and microRNAs: RISCy interactions with serious consequences. *Genes & Development*, 25(18):1881–1894, 2011.
- Czech, B., Malone, C. D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J. A., Sachidanandam, R., et al. An endogenous small interfering RNA pathway in Drosophila. *Nature*, 453:798–802, 2008.
- Davison, A. J., Eberle, R., Ehlers, B., Hayward, G. S., McGeoch, D. J., Minson, A. C., Pellett, P. E., Roizman, B., Studdert, M. J., and Thiry, E. The order herpesvirales. *Archives of virology*, 154(1):171–177, 2009.
- Devroye, L. *Non-Uniform Random Variate Generation*. Springer, New York, 1986.
- Didiano, D. and Hobert, O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature Structural & Molecular Biology*, 13(9):849–851, 2006.
- Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.
- Djuranovic, S., Nahvi, A., and Green, R. A parsimonious model for gene regulation by miRNAs. *Science*, 331(6017):550–553, 2011.
- Djuranovic, S., Nahvi, A., and Green, R. miRNA-Mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science*, 336(6078):237–240, 2012.
- Dölken, L., Malterer, G., Erhard, F., Kothe, S., Friedel, C. C., Suffert, G., Marcinowski, L., Motsch, N., Barth, S., Beitzinger, M., et al. Systematic analysis of viral and cellular microRNA targets in cells latently infected with human gamma-herpesviruses by RISC immunoprecipitation assay. *Cell Host & Microbe*, 7(4):324–334, 2010.

- Dölken, L., Ruzsics, Z., Rdle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P., et al. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, 14(9):1959–1972, 2008.
- Doshi, K. J., Cannone, J. J., Cobaugh, C. W., and Gutell, R. R. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105, 2004.
- Dowell, R. D. and Eddy, S. R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5:71, 2004.
- Dreszer, T. R., Karolchik, D., Zweig, A. S., Hinrichs, A. S., Raney, B. J., Kuhn, R. M., Meyer, L. R., Wong, M., Sloan, C. A., Rosenbloom, K. R., et al. The UCSC genome browser database: extensions and updates 2011. *Nucleic acids research*, 40(Database issue):D918–923, 2012.
- Duursma, A. M., Kedde, M., Schrier, M., le Sage, C., and Agami, R. miR-148 targets human DNMT3b protein coding region. *RNA*, 14(5):872–877, 2008.
- Easow, G., Teleman, A. A., and Cohen, S. M. Isolation of microRNA targets by miRNP immunopurification. *RNA (New York, N.Y.)*, 13(8):1198–1204, 2007.
- Eduati, F., Di Camillo, B., Karbiener, M., Scheideler, M., Corà, D., Caselle, M., and Toffolo, G. Dynamic modeling of miRNA-mediated feed-forward loops. *Journal of computational biology: a journal of computational molecular cell biology*, 19(2):188–199, 2012.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.
- Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836):494–498, 2001.
- Elias, J. E. and Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007.
- Ender, C., Krek, A., Friedländer, M. R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., and Meister, G. A human snoRNA with microRNA-like functions. *Molecular cell*, 32(4):519–528, 2008.
- Erhard, F. *Petri net based system for spatio-temporal simulation of gene regulation*. Diplomarbeit an der Ludwig-Maximilians-Universität München, 2008.
- Erhard, F., Dölken, L., Jaskiewicz, L., and Zimmer, R. PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome Biology*, 14(7):R79, 2013a.

- Erhard, F., Dölken, L., and Zimmer, R. RIP-chip enrichment analysis. *Bioinformatics*, 29:77–83, 2013b.
- Erhard, F., Friedel, C. C., and Zimmer, R. FERN – a java framework for stochastic simulation and evaluation of reaction networks. *BMC Bioinformatics*, 9:356, 2008.
- Erhard, F., Friedel, C. C., and Zimmer, R. FERN - stochastic simulation and evaluation of reaction networks. In S. Choi, editor, *Systems Biology for Signaling Networks*, chapter 30, pages 751–775. Springer, New York, 2010.
- Erhard, F., Haas, J., Malterer, G., Lieber, D., Jaskiewicz, L., Zavolan, M., Dölken, L., and Zimmer, R. Widespread context-dependence of microRNA-mediated regulation. *Genome Research - accepted*, 2013c.
- Erhard, F. and Zimmer, R. Classification of ncRNAs using position and size information in deep sequencing data. *Bioinformatics*, 26(18):i426–432, 2010.
- Erhard, F. and Zimmer, R. Detecting outlier peptides in quantitative high-throughput mass spectrometry data. In *Online Proceedings of German Conference on Bioinformatics*. Weihenstephan, 2011.
- Erhard, F. and Zimmer, R. Detecting outlier peptides in quantitative high-throughput mass spectrometry data. *Journal of Proteomics*, 75:3230–3239, 2012.
- Eulalio, A., Huntzinger, E., and Izaurralde, E. Getting to the root of miRNA-Mediated gene silencing. *Cell*, 132(1):9–14, 2008.
- Evans, T. W., Gillespie, C. S., and Wilkinson, D. J. The sbml discrete stochastic models test suite. *Bioinformatics*, 2007.
- Farazi, T. A., Spitzer, J. I., Morozov, P., and Tuschl, T. miRNAs in human cancer. *The Journal of Pathology*, 223(2):102–115, 2011.
- Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. Potent and specific genetic interference by double-stranded RNA in *caenorhabditis elegans*. *Nature*, 391(6669):806–811, 1998.
- Fisher, S. R. A. *Statistical methods for research workers. Fourteenth Edition Revised*. Oliver & Boyd, 1970. ISBN 0050021702.
- Forshed, J., Johansson, H. J., Pernemalm, M., Branca, R. M. M., Sandberg, A., and Lehtio, J. Enhanced information output from shotgun proteomics data by protein quantification and peptide quality control (PQPQ). *Molecular & Cellular Proteomics*, 2011.
- Fraley, C. and Raftery, A. E. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.
- Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., Leyva, A., Weigel, D., Garcia, J. A., and Paz-Ares, J. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, 39(8):1033–1037, 2007.

- Friedlander, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, 26:407–415, 2008.
- Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2008.
- Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009.
- Frohn, A., Eberl, H. C., Stöhr, J., Glasmacher, E., Rüdell, S., Heissmeyer, V., Mann, M., and Meister, G. Dicer-dependent and -independent argonaute2 protein interaction networks in mammalian cells. *Molecular & Cellular Proteomics*, 11(11):1442–1456, 2012.
- Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., Jovanovich, S. B., Nelson, J. R., Schloss, J. A., Schwartz, D. C., et al. The challenges of sequencing by synthesis. *Nature Biotechnology*, 27(11):1013–1023, 2009.
- Funahashi, A., Tanimura, N., Morohashi, M., and Kitano, H. Celldesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1:159–162, 2003.
- Fundel, K., Haag, J., Gebhard, P., Zimmer, R., and Aigner, T. Normalization strategies for mRNA expression data in cartilage research. *Osteoarthritis and Cartilage*, 16(8):947–955, 2008.
- Gamma, E., Helm, R., Johnson, R., and Vlissides, J. *Entwurfsmuster (German Language)*. Addison-Wesley, 2004.
- Gentleman, R. *Bioinformatics and computational biology solutions using R and Bioconductor*. Birkhäuser, 2005. ISBN 9780387251462.
- Gerard, M. A., Myslinski, E., Chylak, N., Baudrey, S., Krol, A., and Carbon, P. The scaRNA2 is produced by an independent transcription unit and its processing is directed by the encoding region. *Nucleic Acids Res.*, 38:370–381, 2010.
- German, J. B., Hammock, B. D., and Watkins, S. M. Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics: Official journal of the Metabolomic Society*, 1(1):3–9, 2005.
- Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012.
- Ghildiyal, M. and Zamore, P. D. Small silencing RNAs: an expanding universe. *Nature reviews. Genetics*, 10(2):94–108, 2009.
- Gibson, M. and Bruck, J. Efficient exact stochastic simulation of chemical systems with many species and many channels. *Journal of Physical Chemistry A*, 104(9):1876–1889,

- 2000.
- Gillespie, C. S., Wilkinson, D. J., Proctor, C. J., Shanley, D. P., Boys, R. J., and Kirkwood, T. B. L. Tools for the sbml community. *Bioinformatics*, 22(5):628–629, 2006.
- Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.
- Gillespie, D. T. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1–3):404–425, 1992.
- Gillespie, D. T. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.*, 115:1716–1733, 2001.
- Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162:705–708, 1982.
- Gottwein, E., Corcoran, D. L., Mukherjee, N., Skalsky, R. L., Hafner, M., Nusbaum, J. D., Shamulailatpam, P., Love, C. L., Dave, S. S., Tuschl, T., et al. Viral MicroRNA targetome of KSHV-Infected primary effusion lymphoma cell lines. *Cell Host & Microbe*, 10(5):515–526, 2011.
- Goutsias, J. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. *J Chem Phys*, 122(18):184102, 2005.
- Green, E. D. and Guyer, M. S. Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333):204–213, 2011.
- Gresham, D., Dunham, M. J., and Botstein, D. Comparing whole genomes using DNA microarrays. *Nature Reviews Genetics*, 9(4):291–302, 2008.
- Griffiths-Jones, S. The microRNA registry. *Nucleic Acids Research*, 32(suppl 1):D109–D111, 2004.
- Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36:D154–158, 2008.
- Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Molecular Cell*, 27(1):91–105, 2007.
- Grosso, A. R., Martins, S., and Carmo-Fonseca, M. The emerging role of splicing factors in cancer. *EMBO Rep*, 9(11):1087–1093, 2008.
- Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, 2010.
- Gupta, N. and Pevzner, P. A. False discovery rates of protein identifications: A strike against the Two-Peptide rule. *Journal of Proteome Research*, 8(9):4173–4181, 2009.

- Haecker, I., Gay, L. A., Yang, Y., Hu, J., Morse, A. M., McIntyre, L. M., and Renne, R. Ago HITS-CLIP expands understanding of kaposi's sarcoma-associated herpesvirus miRNA function in primary effusion lymphomas. *PLoS Pathog*, 8(8):e1002884, 2012.
- Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano Jr., M., Jungkamp, A.-C., Munschauer, M., et al. Transcriptome-wide identification of RNA-Binding protein and MicroRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.
- Han, J., Lee, Y., Yeom, K., Nam, J., Heo, I., Rhee, J., Sohn, S., Cho, Y., Zhang, B., and Kim, V. Molecular basis for the recognition of primary microRNAs by the drosha-DGCR8 complex. *Cell*, 125(5):887–901, 2006.
- Harris, L. A. and Clancy, P. A "partitioned leaping" approach for multiscale modeling of chemical reaction dynamics. *J Chem Phys*, 125(14):144107, 2006.
- Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A. Z., and Kay, M. A. Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA*, 16(4):673–695, 2010.
- Hausser, J., Syed, A. P., Bilen, B., and Zavolan, M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome research*, 2013.
- He, L. and Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet*, 5(7):522–531, 2004.
- He, L., He, X., Lim, L. P., de Stanchina, E., Xuan, Z., Liang, Y., Xue, W., Zender, L., Magnus, J., Ridzon, D., et al. A microRNA component of the p53 tumour suppressor network. *Nature*, 447(7148):1130–1134, 2007.
- Hendrickson, D. G., Hogan, D. J., Herschlag, D., Ferrell, J. E., and Brown, P. O. Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. *PloS One*, 3(5):e2126, 2008.
- Higgs, P. and Morgan, S. Thermodynamics of RNA folding. when is an RNA molecule in equilibrium? In *Advances in Artificial Life*, pages 852–861. 1995.
- Ho, J. W. K., Bishop, E., Karchenko, P. V., Ngre, N., White, K. P., and Park, P. J. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics*, 12:134, 2011.
- Hobert, O. Gene regulation by transcription factors and MicroRNAs. *Science*, 319(5871):1785–1786, 2008.
- Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. Fast folding and comparison of RNA secondary structures. *Monatshefte fr Chemie / Chemical Monthly*, 125:167–188, 1994.

- Hofacker, I. L., Priwitzer, B., and Stadler, P. F. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20:186–190, 2004.
- Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. COPASI—a COmplex PATHway SIMulator. *Bioinformatics*, 22(24):3067–3074, 2006.
- Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.
- Huttlin, E. L., Jedrychowski, M. P., Elias, J. E., Goswami, T., Rad, R., Beausoleil, S. A., Villn, J., Haas, W., Sowa, M. E., and Gygi, S. P. A Tissue-Specific atlas of mouse protein phosphorylation and expression. *Cell*, 143(7):1174–1189, 2010.
- Ideker, T., Galitski, T., and Hood, L. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2:343–372, 2001.
- Ideker, T. and Lauffenburger, D. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends in biotechnology*, 21(6):255–262, 2003.
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–264, 2003.
- Irizarry, R. A., Wu, Z., and Jaffee, H. A. Comparison of affymetrix GeneChip expression measures. *Bioinformatics (Oxford, England)*, 22(7):789–794, 2006.
- Ivey, K. N. and Srivastava, D. MicroRNAs as regulators of differentiation and cell fate decisions. *Cell Stem Cell*, 7(1):36–41, 2010.
- Jacobsen, A., Wen, J., Marks, D. S., and Krogh, A. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome research*, 20(8):1010–1019, 2010.
- Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Vanario-Alonso, C. E., Samsonova, M., et al. Dynamic control of positional information in the early drosophila embryo. *Nature*, 430(6997):368–371, 2004.
- Jaskiewicz, L., Bilen, B., Hausser, J., and Zavolan, M. Argonaute CLIP - a method to identify in vivo targets of miRNAs. *Methods (San Diego, Calif.)*, 58(2):106–112, 2012.
- Jeang, K.-T. RNAi in the regulation of mammalian viral infections. *BMC Biology*, 10(1):58, 2012.
- Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, 316(5830):1497–1502, 2007.

- Johnston, R. J. and Hobert, O. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature*, 426(6968):845–849, 2003.
- Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. MicroRNAs and their regulatory roles in plants. *Annual review of plant biology*, 57:19–53, 2006.
- Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research*, 7(1):29–34, 2008.
- Kalsotra, A., Wang, K., Li, P.-F., and Cooper, T. A. MicroRNAs coordinate an alternative splicing network during mouse postnatal heart development. *Genes & Development*, 24(7):653–658, 2010.
- Karginov, F. V., Conaco, C., Xuan, Z., Schmidt, B. H., Parker, J. S., Mandel, G., and Hannon, G. J. A biochemical approach to identifying microRNA targets. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19291–19296, 2007.
- Kato, M., de Lencastre, A., Pincus, Z., and Slack, F. J. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. *Genome Biol.*, 10:R54, 2009.
- Kedde, M., Strasser, M. J., Boldajipour, B., Oude Vrielink, J. A. F., Slanchev, K., le Sage, C., Nagel, R., Voorhoeve, P. M., van Duijse, J., Rom, U. A., et al. RNA-binding protein dnd1 inhibits microRNA access to target mRNA. *Cell*, 131(7):1273–1286, 2007.
- Kedde, M., van Kouwenhove, M., Zwart, W., Oude Vrielink, J. A. F., Elkon, R., and Agami, R. A pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nature cell biology*, 12(10):1014–1020, 2010.
- Keller, A., Eng, J., Zhang, N., Li, X.-j., and Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular Systems Biology*, 1:2005.0017, 2005.
- Kennell, D. and Riezman, H. Transcription and translation initiation frequencies of the *Escherichia coli* lac operon. *J Mol Biol*, 114(1):1–21, 1977.
- Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–1284, 2007.
- Khanin, R. and Vinciotti, V. Computational modeling of post-transcriptional gene regulation by microRNAs. *Journal of computational biology: a journal of computational molecular cell biology*, 15(3):305–316, 2008.
- Khanna, A. and Stamm, S. Regulation of alternative splicing by short non-coding nuclear RNAs. *RNA Biology*, 7(4):480–485, 2010.

- Kierzek, A. M. STOCKS: STOChastic Kinetic Simulations of biochemical systems with gillespie algorithm. *Bioinformatics*, 18(3):470–481, 2002.
- Kierzek, A. M., Zaim, J., and Zielenkiewicz, P. The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *J Biol Chem*, 276(11):8165–8172, 2001.
- Kim, H. H., Kuwano, Y., Srikantan, S., Lee, E. K., Martindale, J. L., and Gorospe, M. HuR recruits let-7/RISC to repress c-myc expression. *Genes & development*, 23(15):1743–1748, 2009a.
- Kim, V. N., Han, J., and Siomi, M. C. Biogenesis of small RNAs in animals. *Nature reviews. Molecular cell biology*, 10(2):126–139, 2009b.
- Kincaid, R. P. and Sullivan, C. S. Virus-encoded microRNAs: an overview and a look to the future. *PLoS Pathog*, 8(12):e1003018, 2012.
- Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. A combined computational-experimental approach predicts human microRNA targets. *Genes & development*, 18(10):1165–1178, 2004.
- Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature Methods*, 8(7):559–564, 2011.
- Kitano, H. Computational systems biology. *Nature*, 420(6912):206–210, 2002.
- Knipe, D. M., Howley, P. M., and Griffin, D. E. Fields virology. <http://ebooks.ub.uni-muenchen.de/5638/>, 2007.
- König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7):909–915, 2010.
- Kozak, M. Faulty old ideas about translational regulation paved the way for current confusion about how microRNAs function. *Gene*, 423(2):108–115, 2008.
- Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., et al. Combinatorial microRNA target predictions. *Nature genetics*, 37(5):495–500, 2005.
- Krishnamoorthy, K., Lu, F., and Mathew, T. A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics & Data Analysis*, 51:57315742, 2007.
- Ladewig, E., Okamura, K., Flynt, A. S., Westholm, J. O., and Lai, E. C. Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Research*, 22(9):1634–1645, 2012.

- Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294(5543):853–858, 2001.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.
- Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., et al. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129(7):1401–1414, 2007.
- Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–1831, 2012.
- Landthaler, M., Gaidatzis, D., Rothballer, A., Chen, P. Y., Soll, S. J., Dinic, L., Ojo, T., Hafner, M., Zavolan, M., and Tuschl, T. Molecular characterization of human argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA (New York, N.Y.)*, 14(12):2580–2596, 2008.
- Langenberger, D., Pundhir, S., Ekstrøm, C. T., Stadler, P. F., Hoffmann, S., and Gorodkin, J. deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns. *Bioinformatics (Oxford, England)*, 28(1):17–24, 2012.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.
- Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M., Landthaler, M., and Rajewsky, N. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Molecular cell*, 43(3):340–352, 2011.
- Lee, D.-Y., Zimmer, R., Lee, S. Y., and Park, S. Colored petri net modeling and simulation of signal transduction pathways. *Metab Eng*, 8(2):112–122, 2006.
- Lee, H.-C., Li, L., Gu, W., Xue, Z., Crosthwaite, S. K., Pertsemlidis, A., Lewis, Z. A., Freitag, M., Selker, E. U., Mello, C. C., et al. Diverse pathways generate MicroRNA-like RNAs and dicer-independent small interfering RNAs in fungi. *Molecular Cell*, 38(6):803–814, 2010.
- Lee, R. C., Feinbaum, R. L., and Ambros, V. The *c. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854, 1993.
- Leung, A. K. L., Vyas, S., Rood, J. E., Bhutkar, A., Sharp, P. A., and Chang, P. Poly(ADP-ribose) regulates stress responses and microRNA activity in the cytoplasm. *Molecular*

- cell*, 42(4):489–499, 2011.
- Li, H., Cao, Y., Petzold, L., and Gillespie, D. Algorithms and software for stochastic simulation of biochemical reacting systems. *Biotechnol Prog*, 2007.
- Li, Z., Kim, S. W., Lin, Y., Moore, P. S., Chang, Y., and John, B. Characterization of viral and human RNAs smaller than canonical MicroRNAs. *J. Virol.*, 83:12751–12758, 2009.
- Likić, V. A., McConville, M. J., Lithgow, T., and Bacic, A. Systems biology: The next frontier for bioinformatics. *Advances in Bioinformatics*, 2010:1–10, 2010.
- Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S., and Johnson, J. M. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005.
- Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.
- Linsen, S. E. V., de Wit, E., Janssens, G., Heater, S., Chapman, L., Parkin, R. K., Fritz, B., Wyman, S. K., de Bruijn, E., Voest, E. E., et al. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Meth*, 6(7):474–476, 2009.
- Linsley, P. S., Schelter, J., Burchard, J., Kibukawa, M., Martin, M. M., Bartz, S. R., Johnson, J. M., Cummins, J. M., Raymond, C. K., Dai, H., et al. Transcripts targeted by the MicroRNA-16 family cooperatively regulate cell cycle progression. *Molecular and Cellular Biology*, 27(6):2240–2252, 2007.
- Lipchina, I., Elkabetz, Y., Hafner, M., Sheridan, R., Mihailovic, A., Tuschl, T., Sander, C., Studer, L., and Betel, D. Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. *Genes & Development*, 25(20):2173–2186, 2011.
- Lynch, K. W. Consequences of regulated pre-mRNA splicing in the immune system. *Nat Rev Immunol*, 4(12):931–940, 2004.
- MacRae, I. J., Zhou, K., and Doudna, J. A. Structural determinants of RNA recognition and cleavage by dicer. *Nat Struct Mol Biol*, 14(10):934–940, 2007.
- Makeyev, E. V., Zhang, J., Carrasco, M. A., and Maniatis, T. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Molecular Cell*, 27(3):435–448, 2007.
- Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotech*, 25(1):125–131, 2007.

- Malone, J. H. and Oliver, B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9:34, 2011.
- Marcinowski, L., Tanguy, M., Krmpotic, A., Rädle, B., Lisni, V. J., Tuddenham, L., Chane-Woon-Ming, B., Ruzsics, Z., Erhard, F., Benkartek, C., et al. Degradation of cellular miR-27 by a novel, highly abundant viral transcript is important for efficient virus replication in vivo. *PLoS Pathog*, 8(2):e1002510, 2012.
- Mardis, E. R. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, 2008.
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.
- Maute, R. L., Schneider, C., Sumazin, P., Holmes, A., Califano, A., Basso, K., and Dalla-Favera, R. tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in b cell lymphoma. *Proceedings of the National Academy of Sciences*, 110(4):1404–1409, 2013.
- Meijer, H. A., Kong, Y. W., Lu, W. T., Wilczynska, A., Spriggs, R. V., Robinson, S. W., Godfrey, J. D., Willis, A. E., and Bushell, M. Translational repression and eIF4A2 activity are critical for microRNA-mediated gene regulation. *Science (New York, N.Y.)*, 340(6128):82–85, 2013.
- Mendes, N. D., Freitas, A. T., and Sagot, M.-F. Current tools for the identification of miRNA genes and their targets. *Nucl. Acids Res.*, 37(8):2419–2433, 2009.
- Michalski, A., Cox, J., and Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of Proteome Research*, 10(4):1785–1793, 2011.
- Miller, M. B. and Tang, Y.-W. Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews*, 22(4):611–633, 2009.
- Mishima, Y., Fukao, A., Kishimoto, T., Sakamoto, H., Fujiwara, T., and Inoue, K. Translational inhibition by deadenylation-independent mechanisms is central to microRNA-mediated silencing in zebrafish. *Proceedings of the National Academy of Sciences*, 109(4):1104–1109, 2012.
- Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research*, 18(4):610–621, 2008.
- Morozova, N., Zinovyev, A., Nonne, N., Pritchard, L.-L., Gorban, A. N., and Harel-Bellan, A. Kinetic signatures of microRNA modes of action. *RNA (New York, N.Y.)*, 18(9):1635–1655, 2012.

- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7):621–628, 2008.
- Mukherjee, N., Lager, P. J., Friedersdorf, M. B., Thompson, M. A., and Keene, J. D. Coordinated posttranscriptional mRNA population dynamics during t-cell activation. *Molecular Systems Biology*, 5:288, 2009.
- Mukherji, S., Ebert, M. S., Zheng, G. X. Y., Tsang, J. S., Sharp, P. A., and Oudenaarden, A. v. MicroRNAs can generate thresholds in target gene expression. *Nature Genetics*, 43(9):854–859, 2011.
- Murata, T. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, (77(4)):541–580, 1989.
- Nachmani, D., Lankry, D., Wolf, D. G., and Mandelboim, O. The human cytomegalovirus microRNA miR-UL112 acts synergistically with a cellular microRNA to escape immune elimination. *Nature Immunology*, 11(9):806–813, 2010.
- Naeem, H., Küffner, R., and Zimmer, R. MIRTfnet: analysis of miRNA regulated transcription factors. *PLoS ONE*, 6(8):e22519, 2011.
- Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–1286, 2012a.
- Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012b.
- Nesvizhskii, A. I. and Aebersold, R. Interpretation of shotgun proteomic data. *Molecular & Cellular Proteomics*, 4(10):1419–1440, 2005.
- Nicholson, J. K. and Wilson, I. D. Understanding ‘Global’ systems biology: Metabonomics and the continuum of metabolism. *Nature Reviews Drug Discovery*, 2(8):668–676, 2003.
- Nishikura, K. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nature Reviews. Molecular Cell Biology*, 7(12):919–931, 2006.
- Noble, D. *Music of Life: Biology Beyond Genes*. Oxford University Press, 2008. ISBN 0199228361.
- Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., Gnad, F., Cox, J., Jensen, T. S., Nigg, E. A., et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.*, 3(104), 2010.
- Ong, S.-E. and Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nature chemical biology*, 1(5):252–262, 2005.
- Ozsolak, F. and Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98, 2011.

- Pace, C. N., Heinemann, U., Hahn, U., and Saenger, W. Ribonuclease t1: Structure, function, and stability. *Angewandte Chemie International Edition in English*, 30(4):343360, 1991.
- Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415, 2008.
- Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.
- Park, T., Yi, S., Kang, S., Lee, S., Lee, Y., and Simon, R. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4:33, 2003.
- Pasquinelli, A. E. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics*, 13(4):271–282, 2012.
- Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degen, B., Müller, P., et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, 2000.
- Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.
- Pfeffer, S., Zavolan, M., Grässer, F. A., Chien, M., Russo, J. J., Ju, J., John, B., Enright, A. J., Marks, D., Sander, C., et al. Identification of virus-encoded MicroRNAs. *Science*, 304(5671):734–736, 2004.
- Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.
- Pritchard, C. C., Cheng, H. H., and Tewari, M. MicroRNA profiling: approaches and considerations. *Nature Reviews Genetics*, 13(5):358–369, 2012.
- Puchalka, J. and Kierzek, A. M. Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophys J*, 86(3):1357–1372, 2004.
- Qi, H. H., Ongusaha, P. P., Myllyharju, J., Cheng, D., Pakkanen, O., Shi, Y., Lee, S. W., Peng, J., and Shi, Y. Prolyl 4-hydroxylation regulates argonaute 2 stability. *Nature*, 455(7211):421–424, 2008.
- Raabe, C. A., Hoe, C. H., Randau, G., Brosius, J., Tang, T. H., and Rozhdestvensky, T. S. The rocks and shallows of deep RNA sequencing: Examples in the vibrio cholerae RNome. *RNA*, 17(7):1357–1366, 2011.

- Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology*, 29(5):436–442, 2011.
- Ramsey, S., Orrell, D., and Bolouri, H. Dizzy: stochastic simulation of large-scale genetic regulatory networks. *J Bioinform Comput Biol*, 3(2):415–436, 2005.
- Rao, C. V. and Arkin, A. P. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the gillespie algorithm. *The Journal of Chemical Physics*, 118(11):4999–5010, 2003.
- Rathjen, T., Pais, H., Sweetman, D., Moulton, V., Munsterberg, A., and Dalmay, T. High throughput sequencing of microRNAs in chicken somites. *FEBS Lett.*, 583:1422–1426, 2009.
- Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., and Hatzigeorgiou, A. G. Functional microRNA targets in protein coding sequences. *Bioinformatics (Oxford, England)*, 28(6):771–776, 2012.
- Reddy, V. N., Mavrouniotis, M. L., and Liebman, M. N. Petri net representations in metabolic pathways. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 1:328–336, 1993.
- Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. Genome-wide location and function of DNA binding proteins. *Science (New York, N.Y.)*, 290(5500):2306–2309, 2000.
- Renne, R., Zhong, W., Herndier, B., McGrath, M., Abbey, N., Kedes, D., and Ganem, D. Lytic growth of kaposi’s sarcoma-associated herpesvirus (human herpesvirus 8) in culture. *Nature medicine*, 2(3):342–346, 1996.
- Richard, H., Schulz, M. H., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., et al. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research*, 38(10):e112–e112, 2010.
- Riley, K. J., Rabinowitz, G. S., Yario, T. A., Luna, J. M., Darnell, R. B., and Steitz, J. A. EBV and human microRNAs co-target oncogenic and apoptotic viral and human genes during latency. *EMBO J.*, 31(9):2207–2221, 2012a.
- Riley, K. J., Rabinowitz, G. S., Yario, T. A., Luna, J. M., Darnell, R. B., and Steitz, J. A. EBV and human microRNAs co-target oncogenic and apoptotic viral and human genes during latency. *The EMBO Journal*, 31(9):2207–2221, 2012b.
- Ritchie, W., Flamant, S., and Rasko, J. E. J. Predicting microRNA targets and functions: traps for the unwary. *Nature Methods*, 6(6):397–398, 2009.

- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.
- Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.
- Ruby, J. G., Jan, C. H., and Bartel, D. P. Intronic microRNA precursors that bypass drosha processing. *Nature*, 448(7149):83–86, 2007.
- Rüdel, S., Flatley, A., Weinmann, L., Kremmer, E., and Meister, G. A multifunctional human argonaute2-specific monoclonal antibody. *RNA (New York, N.Y.)*, 14(6):1244–1253, 2008.
- Rüdel, S., Wang, Y., Lenobel, R., Körner, R., Hsiao, H.-H., Urlaub, H., Patel, D., and Meister, G. Phosphorylation of human argonaute proteins affects small RNA binding. *Nucleic acids research*, 39(6):2330–2343, 2011.
- Rybak, A., Fuchs, H., Hadian, K., Smirnova, L., Wulczyn, E. A., Michel, G., Nitsch, R., Krappmann, D., and Wulczyn, F. G. The let-7 target gene mouse lin-41 is a stem cell specific e3 ubiquitin ligase for the miRNA pathway protein ago2. *Nature Cell Biology*, 11(12):1411–1420, 2009.
- Salis, H. and Kaznessis, Y. N. An equation-free probabilistic steady-state approximation: dynamic application to the stochastic simulation of biochemical reaction networks. *J Chem Phys*, 123(21):214106, 2005.
- Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. A ceRNA hypothesis: The rosetta stone of a hidden RNA language? *Cell*, 146(3):353–358, 2011.
- Samant, A., Ogunnaike, B. A., and Vlachos, D. G. A hybrid multiscale monte carlo algorithm (hysmc) to cope with disparity in time scales and species populations in intracellular networks. *BMC Bioinformatics*, 8:175, 2007.
- Samant, A. and Vlachos, D. G. Overcoming stiffness in stochastic simulation stemming from partial equilibrium: a multiscale monte carlo algorithm. *J Chem Phys*, 123(14):144114, 2005.
- Sanger, F. and Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448, 1975.
- Sauer, U., Heinemann, M., and Zamboni, N. Genetics. getting closer to the whole picture. *Science (New York, N.Y.)*, 316(5824):550–551, 2007.
- Sayed, D. and Abdellatif, M. MicroRNAs in development and disease. *Physiological reviews*, 91(3):827–887, 2011.

- Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235):467–470, 1995.
- Schirle, N. T. and MacRae, I. J. The crystal structure of human argonaute2. *Science*, 336(6084):1037–1040, 2012.
- Scholz, M. B., Lo, C.-C., and Chain, P. S. G. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current opinion in biotechnology*, 23(1):9–15, 2012.
- Schwanhaussner, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.
- Schwikowski, B., Uetz, P., and Fields, S. A network of protein-protein interactions in yeast. *Nature biotechnology*, 18(12):1257–1261, 2000.
- Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535–540, 2008.
- Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 2008.
- Sethupathy, P., Megraw, M., and Hatzigeorgiou, A. G. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nature Methods*, 3(11):881–886, 2006.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, 2003.
- Shi, W., Hendrix, D., Levine, M., and Haley, B. A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat. Struct. Mol. Biol.*, 16:183–189, 2009.
- Shimoni, Y., Friedlander, G., Hetzroni, G., Niv, G., Altuvia, S., Biham, O., and Margalit, H. Regulation of gene expression by small non-coding rnas: a quantitative view. *Mol Syst Biol*, 3:138, 2007.
- Shin, C., Nam, J.-W., Farh, K. K.-H., Chiang, H. R., Shkumatava, A., and Bartel, D. P. Expanding the MicroRNA targeting code: Functional sites with centered pairing. *Molecular Cell*, 38(6):789–802, 2010.
- Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. Evolutionarily conserved

- elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.
- Skalsky, R. L., Corcoran, D. L., Gottwein, E., Frank, C. L., Kang, D., Hafner, M., Nusbaum, J. D., Feederle, R., Delecluse, H.-J., Luftig, M. A., et al. The viral and cellular microRNA targetome in lymphoblastoid cell lines. *PLoS pathogens*, 8(1):e1002484, 2012.
- Slepoy, A., Thompson, A. P., and Plimpton, S. J. A constant-time kinetic monte carlo algorithm for simulation of large biochemical reaction networks. *The Journal of chemical physics*, 128(20):205101, 2008.
- Smith, T. F. and Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.
- Speck, S. H. and Ganem, D. Viral latency and its regulation: lessons from the gamma-herpesviruses. *Cell host & microbe*, 8(1):100–115, 2010.
- Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G. J., and Kellis, M. Systematic discovery and characterization of fly microRNAs using 12 drosophila genomes. *Genome Research*, 17(12):1865–1879, 2007.
- Stelling, J. Mathematical models in microbial systems biology. *Current opinion in microbiology*, 7(5):513–518, 2004.
- Stoecklin, G., Tenenbaum, S. A., Mayo, T., Chittur, S. V., George, A. D., Baroni, T. E., Blackshear, P. J., and Anderson, P. Genome-wide analysis identifies interleukin-10 mRNA as target of tristetraprolin. *The Journal of Biological Chemistry*, 283(17):11689–11699, 2008.
- Sturm, M., Hackenberg, M., Langenberger, D., and Frishman, D. TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics*, 11(1):292, 2010.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.
- Szallasi, Z., Stelling, J., and Periwai, V. *System Modeling in Cellular Biology*. MIT Press, 2006.
- Taft, R. J., Glazov, E. A., Lassmann, T., Hayashizaki, Y., Carninci, P., and Mattick, J. S. Small RNAs derived from snoRNAs. *RNA*, 15(7):1233–1240, 2009.
- Tay, Y., Zhang, J., Thomson, A. M., Lim, B., and Rigoutsos, I. MicroRNAs to nanog, oct4 and sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455(7216):1124–1128, 2008.

- Tenenbaum, S. A., Carson, C. C., Lager, P. J., and Keene, J. D. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proceedings of the National Academy of Sciences*, 97(26):14085–14090, 2000.
- Thomas, M., Lieberman, J., and Lal, A. Desperately seeking microRNA targets. *Nat Struct Mol Biol*, 17(10):1169–1174, 2010.
- Thompson, D. M. and Parker, R. Stressing out over tRNA cleavage. *Cell*, 138:215–219, 2009.
- Thomson, D. W., Bracken, C. P., and Goodall, G. J. Experimental strategies for microRNA target identification. *Nucleic Acids Research*, 39(16):6845–6853, 2011.
- Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.
- Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1):46–53, 2013.
- Tu, K., Yu, H., Hua, Y.-J., Li, Y.-Y., Liu, L., Xie, L., and Li, Y.-X. Combinatorial network of primary and secondary microRNA-driven regulatory mechanisms. *Nucleic Acids Research*, 37(18):5969–5980, 2009.
- Umbach, J. L., Nagel, M. A., Cohrs, R. J., Gilden, D. H., and Cullen, B. R. Analysis of human alphaherpesvirus MicroRNA expression in latently infected human trigeminal ganglia. *Journal of Virology*, 83(20):10677–10683, 2009.
- Vasudevan, S. Posttranscriptional upregulation by microRNAs. *Wiley interdisciplinary reviews. RNA*, 3(3):311–330, 2012.
- Vens, C., Rosso, M.-N., and Danchin, E. G. J. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, 27(9):1231–1238, 2011.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. The sequence of the human genome. *Science (New York, N. Y.)*, 291(5507):1304–1351, 2001.
- Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.
- Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research*, 22(9):1680–1688, 2012a.
- Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., et al. Sequence features and chromatin structure around

- the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9):1798–1812, 2012b.
- Wang, Z. and Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA (New York, N.Y.)*, 14(5):802–813, 2008.
- Wang, Z., Gerstein, M., and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009.
- Wee, L., Flores-Jasso, C., Salomon, W., and Zamore, P. Argonaute divides its RNA guide into domains with distinct functions and RNA-Binding properties. *Cell*, 151(5):1055–1067, 2012.
- Westerhoff, H. V. and Palsson, B. O. The evolution of molecular biology into systems biology. *Nature biotechnology*, 22(10):1249–1252, 2004.
- Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S., and Peterson, K. J. The deep evolution of metazoan microRNAs. *Evolution & development*, 11(1):50–68, 2009.
- Wienholds, E. and Plasterk, R. H. A. MicroRNA function in animal development. *FEBS letters*, 579(26):5911–5922, 2005.
- Wightman, B., Ha, I., and Ruvkun, G. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *c. elegans*. *Cell*, 75(5):855–862, 1993.
- Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3:e65, 2007.
- Windhager, L., Bonfert, T., Burger, K., Ruzsics, Z., Krebs, S., Kaufmann, S., Malterer, G., L’Hernault, A., Schilhabel, M., Schreiber, S., et al. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Research*, 2012.
- Witten, D., Tibshirani, R., Gu, S. G., Fire, A., and Lui, W.-O. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biology*, 8(1):58, 2010.
- Witten, J. T. and Ule, J. Understanding splicing regulation through RNA splicing maps. *Trends in Genetics: TIG*, 27(3):89–97, 2011.
- Xu, G., Fewell, C., Taylor, C., Deng, N., Hedges, D., Wang, X., Zhang, K., Lacey, M., Zhang, H., Yin, Q., et al. Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq. *RNA*, 16(8):1610–1622, 2010.
- Yanez-Cuna, J. O., Dinh, H. Q., Kvon, E. Z., Shlyueva, D., and Stark, A. Uncovering cis-regulatory sequence requirements for context specific transcription factor binding.

- Genome Research*, 2012.
- Yang, J.-S., Maurin, T., Robine, N., Rasmussen, K. D., Jeffrey, K. L., Chandwani, R., Papapetrou, E. P., Sadelain, M., O'Carroll, D., and Lai, E. C. Conserved vertebrate mir-451 provides a platform for dicer-independent, ago2-mediated microRNA biogenesis. *Proceedings of the National Academy of Sciences*, 107(34):15163–15168, 2010.
- Yang, X., Zhang, H., and Li, L. Alternative mRNA processing increases the complexity of microRNA-based gene regulation in arabidopsis. *The Plant journal: for cell and molecular biology*, 70(3):421–431, 2012.
- Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.
- Yates, J. R., Eng, J. K., McCormack, A. L., and Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical Chemistry*, 67(8):1426–1436, 1995.
- Zanivan, S., Krueger, M., and Mann, M. In vivo quantitative proteomics: the SILAC mouse. *Methods in molecular biology (Clifton, N.J.)*, 757:435–450, 2012.
- Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., Song, J. J., Kingston, R. E., Borowsky, M., and Lee, J. T. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell*, 40(6):939–953, 2010.
- Zhao, T., Li, G., Mi, S., Li, S., Hannon, G. J., Wang, X.-J., and Qi, Y. A complex system of small RNAs in the unicellular green alga *chlamydomonas reinhardtii*. *Genes & Development*, 21(10):1190–1203, 2007.
- Zhu, L. J., Gazin, C., Lawson, N. D., Pags, H., Lin, S. M., Lapointe, D. S., and Green, M. R. ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, 11:237, 2010.
- Zien, A., Zimmer, R., and Lengauer, T. A simple iterative approach to parameter optimization. *J. Comput. Biol.*, 7:483–501, 2000.

List of abbreviations

nt nucleotides

NGS next generation sequencing

SNP single nucleotide polymorphism

CNV copy number variation

mRNA messenger RNA

ncRNA non-coding RNA

IP immunoprecipitation

4sU 4-thio-uridine

RBP RNA binding protein

FDR false discovery rate

DE differential expression

FDR false discovery rate

LC-MS/MS liquid chromatography tandem mass spectrometry

SILAC stable isotope labelling of amino acids in cell culture

UTR untranslated region

RISC RNA induced silencing complex

GRN gene regulatory network

TF transcription factor

RNAi RNA interference

AGO Argonaute

kb kilobases

EBV Epstein-Barr virus

HSV1 Herpes Simplex virus 1

HSV2 Herpes Simplex virus 2

HCMV Human Cytomegalovirus

MCMV Murine Cytomegalovirus

KSHV Kaposi's Sarcoma-associated Herpesvirus

VZV Varicella-Zoster Virus

rLCV Rhesus monkey Lymphocryptovirus

CLIP crosslinking and immunoprecipitation

PAR-CLIP Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation

eGFP enhanced green fluorescent protein

DGF Digital genomic footprinting

Acknowledgements

Ich möchte mich hier, im letzten Teil meiner Dissertation, bei allen bedanken, die mich in den vergangenen Jahren unterstützt und mir in vielerlei Hinsicht geholfen haben.

Besonderen Dank schulde ich Prof. Dr. Ralf Zimmer, der mir die Möglichkeit und jede Freiheit gegeben hat, an so vielfältigen und spannenden Themen zu arbeiten, der mich immer zu eigenen Ideen ermuntert hat und der immer bereit zu vielen ergiebigen Diskussionen war. Ich möchte mich bei ausßerdem all meinen Kollegen bedanken, nicht nur für die tolle Zusammenarbeit, sondern ganz besonders auch für alles außerhalb der Bioinformatik was wir zusammen erleben durften.

Mein Dank gilt auch unseren Kooperationspartnern, allen voran Dr. Lars Dölken und Prof. Dr. Dr. Jürgen Haas, deren experimentelle Arbeiten mir erst ermöglicht haben, all diesen interessanten Problemstellungen nachzugehen und von denen ich dank häufiger und interessanter Diskussionen viel gelernt habe.

Ich möchte mich auch bedanken bei Prof. Dr. Rolf Backofen und Prof. Dr. Hans Jürgen Ohlbach, die sich dazu bereit erklärt haben, meine Promotion als Zweitgutachter und als Prüfungskommissionsvorsitzender zu unterstützen.

Ich bin auch unendlich dankbar gegenüber meiner Familie und meinen Freunden, für all die Unterstützung und vor allem Ablenkung in den Jahren dieser Arbeit und auch schon zuvor.

Die letzten geschriebenen Zeilen dieser Dissertation möchte ich meiner Frau Sandra widmen. Es ist bestimmt manchmal nicht einfach, wenn der Ehemann einen so unverständlichen Beruf hat. Das geht mit Gesprächen los wie “Was macht denn Ihr Mann beruflich?” “Aha, ist ja interessant, und was macht man so als Bioinformatiker?”, und endet sicherlich nicht mit “Schatz, wie war Dein Tag heute?” “Naja, das neue Modell, dass ich für die PAR-CLIP-Daten entwickelt habe, hatte leider nicht so große Auswirkungen auf die ROC Kurve wie ich erwartet hab, und dabei berücksichtigt es doch neben den Thymin-zu-Cytosin Konversionen jetzt auch den Bias durch die RNase T1, die ja bekanntlich nach Guanin schneidet”. Trotzdem, in dieser Arbeit, die über die letzten Jahre entstanden ist, steckt auch sehr viel von Dir, Sandra, denn ohne die Kraft, die Du mir immer gibst, ohne die Leichtigkeit, mit der ich daheim auch mal von der Arbeit abschalten kann, und ohne Dein Verständnis, wenn ich dank wichtiger deadlines mal wieder länger arbeite, hätte ich diese Arbeit so nicht schreiben können.